THESIS

FOR THE DEGREE OF Ph.D.

"ON THE THEORY OF ESTIMATION OF STATISTICAL PARAMETERS"

Harold Silverstone, M.A. (N.Z.)

May, 1939.

CHAPTER ONE.

INTRODUCTION.

1·00 The problem of estimation has figured prominently in the mathematical theory of statistics during the last eighteen or twenty years, although, of course, it has been one of the central problems of statistics since its inception as a separate branch of mathematical science. In recent years the study of the problem has been given a most powerful stimulus through the investigations of R. A. Fisher, who by a special combination of practical and theoretical insight has realised vividly the inner nature of the problem and the inductive nature of the reasoning processes employed in its solution.

1·01 The Nature of the Problem.

From previous experience, or from actual observation of a sufficiently large sample of measurements, we are usually able to say something concerning the mathematical form of the population from which the sample has been drawn - for instance we can tell whether the population is a normal population, or one following Pearson's type III law, and so on. The problem of deciding the mathematical form of the population is known as the problem of "specification"; and the problem of estimation starts from the point where the problem of specification has been solved. The problem of estimation may

be stated as follows:

Given a sample of n observations, $x_1, x_2, \cdots, x_n$, drawn from a population whose mathematical form (orprobability differential) is

$$dp = \varphi(x \mid \theta_1, \theta_2 \cdots, \theta_n)\, dx$$

what functions, $T_1(x_1, \cdots, x_n)$, $T_2(x_1, \cdots, x_n), \cdots$ $\cdots T_n(x_1, \cdots x_n)$, will give an adequate estimate of the unknown parameters, $\theta_1, \theta_2, \cdots, \theta_n$, which serve fully to specify the population ? The x_i and x can be either scalar or vector quantities. For example, for two variables x and y we have n pairs of observations (x_i, y_i) .

1·02 Deduction and Induction.

It is obvious that a problem of this type will differ fundamentally from the common deductive problems in the theory of probability. Such a deductive problem might, for instance, be:

Given the population φ, where the values of $\theta_1, \cdots, \theta_n$ are known, to find the probability of occurrence of any particular sample of n . Such a problem presents no difficulties at all. The difficulty of the problem of estimation rises from the fact that we are arguing not from the population to the sample, but from the sample to the population - from the particular to the general. The former

is the deductive method - the latter is the inductive method. Of course the above division between deduction and induction is only a broad division applicable to the whole process but not to every part of the process. There is no contradiction involved in the statement that each step in the reasoning process of induction is at the same time a deductive step. It would be impossible to conduct any process of reasoning from the particular to the general without the use of deductive reasoning at each step in the process. In the same way every deductive process, whilst preserving as a whole its character as a deductive process, involves inductive reasoning at each particular step in that process. However, despite this interrelation between deduction and induction, problems of induction must be distinguished from problems of deduction, and must receive special treatment. Moreover it will be incorrect to judge the validity of an inductive process exclusively by the standards of judgement set by deductive processes.

1·03 The Method of Solution.

Having thus perceived the special nature of our problem, let us see how it can be solved. In the first place we must realise that we have no information concerning the values of the parameters except that supplied by the sample itself. In other words we have no prior knowledge concerning the parameters, and to assume that we have prior knowledge, to assume, for

instance, that the parameters have a certain prior distribution, not only is unwarranted but destroys the whole basis of our problem and its nature as a problem in induction.

On the other hand, to make assumptions, concerning the nature of our estimating functions, T, is not only justified but is necessary. Such assumptions are a central factor in the solution of the problem, (as we shall see later), and are connected with obtaining what we have termed "adequate" functions from which to estimate the values of $\theta_1, \cdots, \theta_k$. Again it is obvious that if a number of functions present themselves as being "adequate" for the estimation of a particular parameter θ, we shall be obliged to make a choice of one particular function, which we shall call the "best" estimate of θ. The notion of the "best" estimate will depend upon the satisfaction of certain criteria, or, on the other hand, might be determined by the purpose for which we desire to use the estimate.

1·04 The Method of Moments.

During the first quarter of the present century it might perhaps have been said that the most popular method of estimation was the "method of moments" due to Karl Pearson. By this method, if we wish to estimate k parameters, $\theta_1, \cdots, \theta_k$, specifying the population

$$\varphi(x \mid \theta_1, \cdots, \theta_k),$$

we equate the first k theoretical moments of the curve φ

to the first K moments of the observed sample. We thus obtain K equations from which to solve for the K unknowns.

For instance the problem of estimating a and c in a frequency curve of the form

$$\varphi\left(\frac{x-a}{c}\right),$$

i.e. the problem of <u>locating</u> and <u>scaling</u> the curve, is solved according to the method of moments by the use of the first and second moments in the above manner.

But it happens that there are curves for which the first and second moments are of very little use or of no use at all.

If we take, for example, the so-called Cauchy distribution

$$d\varphi = \frac{1}{\pi}\frac{dx}{1+(x-a)^2},$$

we find that its second moment is infinite, and so is quite useless.

Moreover, if we consider

$$d\varphi = \frac{a}{\pi}\frac{dz}{a^2+z^2},$$

and find the sampling distribution of the mean, we have:

The moment generating or characteristic function of the curve is

$$\varphi_1(t) = \frac{a}{\pi}\int_{-\infty}^{\infty} e^{itz}\frac{dz}{a^2+z^2}.$$

A pole is $z = ia$. If we take a contour encircling the pole and formed by the real axis and a semicircle of

of infinite radius, we get e^{-at} from the residue. [See P. Lévy, "Calcul des Probabilités" (1925), p.p.179-180.] As the function is even, we have

$$\varphi_1(t) = e^{-a|t|}$$

Hence the moment generating function of

$$\frac{1}{\pi} \frac{1}{1+(x-a)^2}$$

is

$$e^{ita-|t|},$$

so that the characteristic function of the mean of n sample values is

$$e^{n(ita/n - |t|/n)},$$

i.e.

$$e^{ita-|t|},$$

the same as that of the population. (Exactly the same result can readily be seen to hold for any weighted mean with positive weights).

Hence the mean of a sample of n observations has exactly the same distribution as has a single observation, and so is liable to errors of the same magnitude. Accordingly the mean is no better in locating the unknown origin $x=a$ than any

single observation.

On the other hand the median of this curve has a standard error tending in large samples to $\pi/(2\sqrt{n})$. Hence the median is much superior to the mean for locating the Cauchy curve.

R. A. Fisher, ["On the Mathematical Foundations of Theoretical Statistics", Phil. Trans., A 222, pp 309-368, (1921)], has shown that the only curve of the Pearsonian types for which the mean is the best estimate of a is the normal curve.

1·05 "Presumptive" Moments.

An early problem in the theory of estimation was that of finding the "best" estimate of the population variance for any population, when we have a sample of n from that population. The moment so obtained was termed a "presumptive moment".

Let

$$m_r = \frac{1}{n}\sum_i x_i^r \quad , \quad \mu_r = \text{population value},$$

$$s_r = \frac{1}{n}\sum_i (x_i - m_1)^r \quad , \quad \sigma_r = \text{population value}.$$

As n increases indefinitely, $m_r \rightsquigarrow \mu_r$, and $s_r \rightsquigarrow \sigma_r$.

When n is finite, what are the "best" estimates of μ_2 and σ_2 ? Let these estimates be $\hat{\mu}_2$ and $\hat{\sigma}_2$.

Gauss solved the problem as follows:

$$s_2 = m_2 - m_1^2$$

$$= \frac{1}{n} \Sigma x_i^2 - \frac{1}{n^2} (\Sigma x_i)^2$$

$$= \frac{1}{n}\left(1 - \frac{1}{n}\right) \Sigma x_i^2 - \frac{2}{n^2} \Sigma x_i x_j \quad , \; i \neq j ,$$

$$\therefore \; E(s_2) = \frac{1}{n}\left(1 - \frac{1}{n}\right) \Sigma E(x_i^2) - \frac{2}{n^2} \Sigma E(x_i) E(x_j)$$

since x_i and x_j are independent.

$$\therefore \; E(s_2) = \frac{1}{n}\left(1 - \frac{1}{n}\right) n \mu_2 - \frac{2}{n^2} \binom{n}{2} \mu_1^2$$

$$= \frac{n-1}{n} (\mu_2 - \mu_1^2)$$

$$= \frac{n-1}{n} \sigma_2 .$$

This is an exact formula. But as we have only one sample, we must take

$$E(s_2) = s_2$$

for obtaining $\hat{\sigma}_2$.

Hence

$$\hat{\sigma}_2 = \frac{n}{n-1} s_2 . \qquad \cdots \quad (1)$$

In criticism of this, Steffensen, ["Some Recent Researches in the Theory of Statistics and Actuarial Science" Cambridge, (1930) ,] puts forward the following arguments.

By the same method, we have

$$E(m_2) = \frac{1}{n} \Sigma E(x_i^2) = \mu_2$$

$$\therefore \; \hat{\mu}_2 = m_2 \qquad \cdots \quad (2)$$

Steffensen comments upon the "very suspicious fact that, whereas μ_2 and σ_2 must be considered "equally good" (for what purpose he does not say), yet the application of the same method leads to the use of m_2 unaltered for $\hat{\mu}_2$ but to $\frac{n}{n-1} s_2$ for $\hat{\sigma}_2$.

Moreover, according to Steffensen, since

$$\sigma_2 = \mu_2 - \mu_1^2 , \quad \ldots \quad (3)$$

we must have

$$\hat{\sigma}_2 = \hat{\mu}_2 - \hat{\mu}_1^2 , \quad \ldots \quad (4)$$

i.e.
$$\frac{n}{n-1} s_2 = m_2 - m_1^2 = s_2 ,$$

$$\therefore \quad n = n-1$$

which is absurd.

But, in truth, there is nothing "suspicious" in the difference between forms (1) and (2)

In the first place, (4) can only be inferred from (3) as an exact relation, when n is infinite; and in that case $n=n-1$.

Secondly, when n is finite, the discrepancy arises from the fact that we have replaced $E(s_2)$, $E(m_2)$ and $E(m_1)$ by s_2, m_2 and m_1. But the non-linear functional relationship connecting s_2, m_2 and m_1, (viz. $s_2 = m_2 - m_1^2$), will not be the functional relationship connecting $E(s_2)$ $E(m_2)$ and $E(m_1)$; and so, whereas equation (3) is correct, equation (4) is not correct.

The discrepancies that thus arise are certainly important enough when n is small, but, from what has preceded, it is easy to see that they are relatively unimportant in comparison with the standard sampling error due to the nature of the probability curve.

Moreover these so-called "presumptive" moments are not of much use to us if nothing is known or assumed concerning the curve.

1·06 Locating and scaling the Normal Curve.

$$\phi = \frac{1}{\sqrt{2\pi}\cdot\sigma}\, e^{-\frac{1}{2}(x-a)^2/\sigma^2}.$$

We have to estimate a and σ

To estimate a

Consider $m_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i)$

The distribution of m_1 in samples of n is

$$dp = \frac{1}{\sqrt{2\pi n}\,\sigma}\, e^{-\frac{1}{2}(m-a)^2/n\sigma^2}\, dm, \quad \cdots (1)$$

a normal curve with mean, median and mode all coinciding at a.

It can be shown that m_1 is the best functional estimate of a.

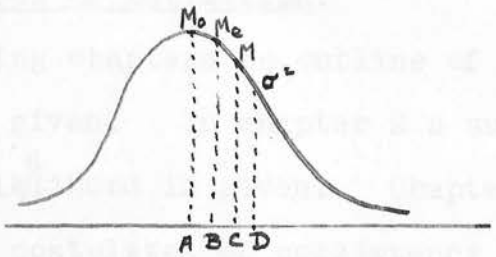
Now the m_1 that we have from our sample might be regarded as a single observation from a population of "m_1's", distribut ed according to (1). The problem is, "where shall we place it ?" If we take it as being typical of the population of m_1's we would place it at the centre of the distribution, where the mean, the median and the mode coincide. Having thus located the curve by means of m_1, we have yet to append a standard error $\sigma/\sqrt{n}$; and this involves the estimation of σ.

To estimate σ

The second moment v, as calculated from the sample, in the usual manner, has a sampling distribution given by

$$dp = c\, e^{-\frac{1}{2}nv/\sigma^2}\, v^{\frac{1}{2}(n-3)}\, dv, \qquad \cdots (2)$$

a type III, or Gamma type, distribution with mean $\frac{n-1}{n}\sigma^2$; mode $\frac{n-3}{n}\sigma^2$; and median approximately $\frac{3n-5}{3n}\sigma^2$, unless n is very small and the curve consequently very skew. The relative positions of the mean M, the median M_e, and the mode M_o are shown below.



Again our second moment <u>calculated from the sample</u> is to be taken as typical and we are faced with the problem of where to place it on the curve. The convention by which $n-1$ is used as divisor instead of n places it at C. Steffensen places it at D; i.e. he locates the variance curve with extreme bias, and does not appear to have good reason for doing so. (Steffensen merely says it is "safer" to put $\hat{\sigma}_2 = s_2$).

Our choice would probably be between the mean and the median, as these have useful mathematical properties. The mean is that value about which the sum of squared deviations

is a minimum; and the median that value about which the sum of absolute deviations is a minimum. The median also has the useful property of invariance under a monotonic functional transformation. An examination of the consequences of this process will be found in the chapter on the "empirical methods".

The mode is not worthy of much consideration as its mathematical properties are not so useful - and the fact that the mode is the "most probable" value does not mean that it is a "very probable" value.

1·07 The Methods of Estimation.

In the following chapters an outline of the chief methods of estimation are given. In chapter 2 a summary of Fisher's work on maximum likelihood is given. Chapter 3 sets out a new method based upon postulates of consistency and minimum sampling error. Under chapter 4 are grouped the various "empirical" methods such as we have considered in this chapter. Chapter 5 contains a brief account of the methods of estimation by interval, due mainly to the work of J. Neyman and E. J. G. Pitman.

The main point that will emerge from this examination is that none of the methods is universal in its application.

We have seen already where the method of moments breaks down.

If we try to apply maximum likelihood to locating and scaling the curve

$$\varphi = \frac{1}{c} e^{-(x-a)/c}$$

We will find $a = -\infty$, which is useless. If we take a as the mean $\bar{x}$ of the sample we have

$$\frac{\partial}{\partial c} \log \left\{ c^{-n} e^{n(a-\bar{x})/c} \right\} = 0,$$

i.e.

$$\frac{\partial}{\partial c} \left\{ -n \log c + n(a-\bar{x})/c \right\} = 0,$$

i.e.

$$c = \bar{x} - a = 0,$$

which, again, is useless.

Further we will find that in many cases the method of maximum liklihood leads to great practical difficulties. For instance the maximum liklihood estimate of a in the Cauchy distribution leads to the necessity of solving an equation of degree $2n+1$ in a.

In such a case we might feel justified in abandoning the "best" estimate and using one that is "almost as good" provided we can supply a standard sampling error.

For instance a good estimate for θ in the rectangular distribution $\frac{dx}{\theta}$, is the mean of the extreme observations. Now the Cauchy distribution,

$$dp = \frac{1}{\pi} \frac{dx}{1+(x-a)^2}$$

is the distribution of the points of intersection with a fixed line of a random ray drawn through a fixed point at unit distance from the line. By considering not these points, but the distribution of θ, the angle between the ray and a fixed line through the given point, we replace the Cauchy distribution by a rectangular distribution. A good estimate for θ is then

seen to be

$$\tan\left\{\tfrac{1}{2}(\arctan x_1 + \arctan x_n)\right\}$$

where x_1 and x_n are the extreme observations.

Difficulties of practical application also attend the methods of estimation by interval, whilst a further difficulty is often created by a lack of uniqueness in the solution, which seriously limits the number of distributions for which the method can be used with advantage.

CHAPTER TWO

The Method of Maximum Likelihood.

2·00 It has frequently been pointed out, especially by critics of the work of R. A. Fisher, that the priority in discovering the method of maximum likelihood belongs not to R. A. Fisher but to much earlier writers, most notable amongst whom was Gauss himself. However this may be, (and Fisher does not feel himself called upon to "refute" such an assertion,) two things are certain: firstly, that Fisher "turned Gauss right side up," and secondly that by so doing Fisher freed Gauss's method from the bonds of purely deductive reasoning and gave to it a scope and general validity not conceived of by Gauss himself.

2·01 Our problem is as follows: given a sample of n observations $x_1, x_2, \cdots\cdots, x_n$, where x_i is distributed according to

$$\phi(x_i | \theta_1, \cdots\cdots, \theta_k)\, dx_i ,$$

we require to find the "best" estimates of the θ_j, basing ourselves (1) on the observed values x_i; (2) on the assumed form of the population.

It is obvious that there can be any number of functions of the x_i which could serve to tell us something about the values of the θ_j. Any such function is called an "estimate" of the appropriate θ_j, or, in Fisher's terminology a

"statistic" for the estimation of θ_j .

2·02. In order to limit the number of statistics, and finally to arrive at a certain "best" statistic, it is necessary to introduce a number of restrictive conditions to be imposed on our estimate. One obvious condition that must apply, no matter what our ultimate method may be is the condition or criterion of "consistency", which states that, in the limit, when our sample tends in size to the population, our estimating function must yield θ_j itself. Confining our attention for the moment to a single parameter θ , we may state this condition more precisely as follows:

If $T(x_1, \ldots, x_n)$ is a statistic used for the estimation of θ , given a sample of n , then the probability that

$$|T(x_1, \ldots, x_n) - \theta| < \varepsilon$$

tends to unity as n tends to infinity, where ε is a positive quantity no matter how small. Again, it is obvious that this condition alone will not serve uniquely to define T but leaves us the problem of choosing between various statistics satisfying the criterion of consistency.

2·03 The problem now is: given various consistent statistics $T_1, T_2, \ldots$, which may be used to estimate θ , which one shall we choose ? It is quite natural to suggest one possible answer, namely : we will choose the "most reliable" one in the sense that if we take samples of n in all possible

ways from the same population and find the respective values of $T_1, T_2, \cdots\cdots$ for each sample, we will choose that which has the smallest variance.

This condition gives what Fisher calls the criterion of "efficiency," and may be stated thus: "The fixed value to which the variance of a statistic (multiplied by n) tends shall be as small as possible."

If T is such a statistic and T_1 any other statistic, and if σ_T^2 , $\sigma_{T_1}^2$ are their respective variances when n is large, then we have in the ratio $\frac{\sigma_T^2}{\sigma_{T_1}^2}$ a measure of the relative efficiency of T_1 , the efficiency of T, since σ_T^2 is a minimum, being taken as 100%.

A statistic T that is 100% efficient is said to use <u>all</u> the information available in the sample when the sample is large. The percentage of the information used by any other statistic T_1 is

$$\frac{\sigma_T^2}{\sigma_{T_1}^2} \times 100\% .$$

For example, in the case of the normal curve it may be shown that the mean is a 100% efficient statistic (i.e. is a minimum variance estimate) and has variance $\frac{\sigma^2}{n}$. The median, on the other hand, has a variance tending, when n is large to $\frac{\pi}{2n}\sigma^2$. Hence the relative efficiency of the median is

$$\frac{\sigma^2}{n} \Big/ \frac{\pi\sigma^2}{2n} = \frac{2}{\pi} = 63{\cdot}66\%$$

i.e. the median utilises only 63.66% of the information available in the sample.

2·04. This notion of identifying the relative efficiency of T_1 with the amount of information used by T_1 has been adversely criticised as being unclear and quite arbitrary. However its justification appears in the following:

If e = relative efficiency of T_1,

then $$\frac{n\sigma_T^2}{n\sigma_{T_1}^2} = e \quad , e<1$$
$$= \frac{V_T}{V_{T_1}} \quad , \text{ say.}$$

Hence $\sigma_T^2 = V_T/n$

and $\sigma_{T_1}^2 = V_T/(en)$

So, to make $\sigma_{T_1}^2 = \sigma_T^2$, we would have to replace n in the second case by n' where

$$en' = n$$
$$\text{or} \quad n' = \frac{n}{e}$$

that is we would have to increase the size of our sample in the ratio $\frac{1}{e}$ if we are to use T_1 in such a way as to be equally reliable as T.

It is in this sense that we may say that in using T_1 in a sample of n we are rejecting a certain fraction of the information.

One drawback to the above conceptions is that it is possible that the relative efficiencies of different statistics may tend to equality when n is large, yet be different for

small samples. Hence it will be necessary to find a definition for efficiency in finite samples.

In the meantime, however, let us return to our problem of finding the best estimate of θ .

2·05 Given our two conditions (of consistency and efficiency) we might proceed in either of the two following ways to find a unique estimate T .

(1) We might replace our consistency condition by one that is more rigorous and which will, when combined with the efficiency condition, yield a unique estimate.

(2) We might introduce a third condition that will yield a unique estimate, and proceed to show that this estimate satisfies the two conditions of consistency and efficiency.

It is the second method that Fisher pursues and for the purpose of which he introduces the notion of maximum likelihood.

Later it will be shown that it is possible to arrive at substantially the same result as Fisher's by following the first course.

2·06 As fundamental notions tend to be obscured through the familiarity induced by constant application of a method, it will be worth while to quote the precise words in which Fisher outlined his ideas in his 1921 paper.

On page 323 we find the following:

"If in any distribution involving unknown parameters $\theta_1, \theta_2, \theta_3, \text{----}$, the chance of an observation falling in the range dx be represented by

$$f(x \mid \theta_1, \theta_2, \text{----})\, dx$$

then the chance that in a sample of n, n_1 fall in the range dx_1, n_2 in the range dx_2, and so on, will be

$$\frac{n!}{\prod(n_p!)} \prod \left\{ f(x_p \mid \theta_1, \theta_2, \text{-----})\, dx_p \right\}^{n_p}$$

The method of maximum lik/e/lihood consists simply in choosing that set of values for the parameters which makes this quantity a maximum, and since in this expression the parameters are only involved in the function f, we have to make

$$S(\log f)$$

a maximum for variations of $\theta_1, \theta_2, \theta_3$, etc."

For a continuous distribution where the probability of an observation x falling within the range dx is

$$\varphi(x \mid \theta_1, \text{---}, \theta_n)\, dx$$

the compound probability of our sample $x_1, \text{-----}, x_n$ occurring is

$$\Phi(x_1, \text{---}, x_n \mid \theta_1, \text{---} \theta_n)\, dx_1 \text{----} dx_n$$
$$= \prod_{i=1}^{n} \left\{ \varphi(x_i \mid \theta_1, \text{---}, \theta_n) \right\} dx_1 \text{---} dx_n$$

What we maximise is not the probability $\Phi(x \mid \theta)\, dx$, but the "lik/e/lihood", $\Phi(x \mid \theta)$, in fact the "probability

density". The "best" value of θ is taken to be that value which gives the greatest liklihood of our actual sample occurring.

2·07 As probability is really a relative frequency we can speak of the probability of our sample occurring when the population is fully specified, but given a sample alone, we can only speak of the "likelihood" of a state of affairs existing in the population. Of course the word "likelihood" has no magic properties in itself, but serves as a convenient term to distinguish an inductive reasoning process from a deductive one. The actual fact appears to be that we are concerned with a probability density and not with a probability; to speak of maximising

$$\Phi(x|\theta)\,dx$$

would not seem to have much meaning, since $\Phi(x|\theta)\,dx$ is an infinitesmally small quantity.

2·08 The Normal Curve

$$\phi(x|m,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}(x-m)^2/\sigma^2}$$

$$\log \Phi = \text{const} - n\log\sigma - \tfrac{1}{2}\Sigma(x_r-m)^2/\sigma^2$$

$$\frac{\partial}{\partial m}\log\Phi = \frac{1}{\sigma^2}(\Sigma x_r - nm).$$

∴ best value, $\hat{m}$, of m is given by

$$\hat{m} = \frac{1}{n}\Sigma x_r = \bar{x}.$$

$\frac{\partial}{\partial\theta}\log\Phi = 0$ yields

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma(x_r-m)^2, \quad \text{when } \theta=\sigma,$$

and not $\frac{1}{n-1}\Sigma(x_r-m)^2$. On the other hand m may be unknown.

2·09. The curve $c e^{-\frac{1}{2}|x-\theta|}$

$$\Phi = c^n e^{-\frac{1}{2}\sum_1^n |x_r-\theta|}$$

The maximum liklihood solution is that value of θ which maximises Φ ; *i.e.* is that value which minimises $\Sigma|x_r-\theta|$; *i.e.* is the median of the x_r .

2·10 The curve $c e^{-(x-m)}$

$$\log \Phi = -(\Sigma x_r - nm)$$

$$\therefore \frac{\partial}{\partial m} \log \Phi = n .$$

Hence no maximum likelihood solution exists.

2·11 The rectangular distribution of known centre but unknown range.

$$\phi(x|\theta)\,dx = \frac{dx}{\theta}$$

$$\Phi = \theta^{-n}$$

Again no maximum likelihood solution exists.

2·12 The Cauchy distribution of unknown centre.

$$\phi(x|\theta)\,dx = \frac{1}{\pi}\,\frac{dx}{1+(x-\theta)^2}$$

$$\frac{\partial}{\partial\theta}\log\Phi = \Sigma\left\{\frac{\partial}{\partial\theta}\log\phi\right\}$$

$$= \text{const.} + \Sigma\left\{\frac{2(x_r-\theta)}{1+(x_r-\theta)^2}\right\},$$

leaving

$$\sum_{r=1}^{n} \frac{x_r-\hat{\theta}}{1+(x_r-\hat{\theta})^2} = 0$$

to be solved as an equation in $\hat{\theta}$, of degree $2n+1$.

For two observations, x_1 and x_2, the solution is

$$\hat{\theta} = \tfrac{1}{2}(x_1 + x_2)$$

For the unlikely case of $2n+1$ observations equally spaced the solution is

$$\hat{\theta} = x_n$$

For $2n$ observations equally spaced the solution is

$$\hat{\theta} = \tfrac{1}{2}(x_n + x_{n+1}).$$

For other systems of $x_1, \cdots, x_n$ we must solve the equations by approximate means. For instance we might find by trial and error a value of $\theta_1 (= \theta_1)$, which makes $\Sigma \frac{x_r - \theta}{1+(x_r-\theta)^2}$, nearly zero, then find the values of $\Sigma \frac{x_r - \theta}{1+(x_r-\theta)^2}$ for two values of θ in the neighbourhood of θ_1 and obtain $\hat{\theta}$ by parabolic interpolation.

Although the equation

$$\Sigma \frac{x_r - \theta}{1+(x_r-\theta)^2} = 0$$

has $2n+1$ roots, only one of them will lie in the range $(x_1, \cdots x_n)$ where x_1, x_n are the smallest and largest observations respectively. It is this root that we must take for $\hat{\theta}$. Further, since all the x_r are positive there can be no real root lying outside the range $(x_1, \cdots, x_n)$ Hence the solution is unique.

2·13. To establish the validity of the method of maximum likelihood we have to show that the solution of the equation

$$\frac{\partial}{\partial \theta} \log \Phi$$

yields a statistic $\hat{\theta}$ satisfying the criteria of consistency and efficiency.

2·14 The Consistency of $\hat{\theta}$

$\hat{\theta}$ is the solution of

$$\frac{\partial}{\partial\theta}\log\Phi(x|\theta)=0$$

Now

$$\log\Phi(x|\theta)=\Sigma\log\phi(x|\theta)$$

where $\phi(x|\theta)$ is the elementary probability law.

Hence $\frac{1}{n}\log\Phi(x|\theta)=\frac{1}{n}\Sigma\log\phi(x|\theta)$, and so when $n\rightsquigarrow\infty$, $\frac{1}{n}\frac{\partial}{\partial\theta}\log\Phi(x|\theta)$ tends to the expectation of $\frac{\partial}{\partial\theta}\log\phi(x|\theta)$; i.e. tends to

$$\int\phi(x|\theta)\frac{\partial}{\partial\theta}\log\phi(x|\theta)\,dx$$

$$=\int\phi\cdot\phi_\theta/\phi\,dx=\int\phi_\theta\,dx=0$$

since

$$\int\phi\,dx=1$$

The condition is that differentiation under the integral sign is permissible.

Now $\hat{\theta}$ satisfies

$$\frac{1}{n}\frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta})=0\quad,\text{ whilst }\frac{1}{n}\frac{\partial}{\partial\theta}\log\Phi(x|\theta)\rightsquigarrow 0.$$

Hence if $\left|\frac{1}{n}\frac{\partial}{\partial\theta}\log\Phi(x|\theta)-\frac{1}{n}\frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta})\right|<\varepsilon$, it follows that $\left|\frac{\partial}{\partial\theta}\log\phi(x|\theta)-\frac{\partial}{\partial\theta}\log\phi(x|\hat{\theta})\right|<\varepsilon$;

and so $|\theta-\hat{\theta}|<\varepsilon'$;

and $\hat{\theta}$ is a consistent statistic.

2·15 The efficiency of $\hat{\theta}$.

To prove the efficiency of $\hat{\theta}$, it is necessary first to find an expression for the variance of $\hat{\theta}$, and then to show that the variance of any other estimate, θ_1, say, is greater than that of $\hat{\theta}$.

2·16 The variance of $\hat{\theta}$.

We may write $\frac{\partial}{\partial\theta}\log\Phi(x|\theta)$ in the form

$$\frac{\partial}{\partial\theta}\log\Phi(x|\theta) = \frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) + (\theta-\hat{\theta})\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\hat{\theta}) + \frac{1}{2!}(\theta-\hat{\theta})^2\frac{\partial^3}{\partial\theta^3}\log\Phi(x|\hat{\theta}) + \cdots$$

or

$$\frac{\partial}{\partial\theta}\log\Phi(x|\theta) = \frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) + (\theta-\hat{\theta})\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\theta'),$$

where θ' lies between θ and $\hat{\theta}$.

But $\frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) = 0$; and when n is large, it follows, due to the consistency of $\hat{\theta}$, that

$\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\theta')$, $= \frac{\partial^2}{\partial\theta^2}\Sigma\{\log\varphi(x|\theta')\}$, converges to $n E\{\frac{\partial^2}{\partial\theta^2}\log\varphi(x|\theta)\}$, $= n\bar{b}$ say.

Hence if $E' \equiv$ expectation over all samples of n, we have

$$E'(\theta-\hat{\theta})^2 = \frac{1}{n^2\bar{b}^2}E'\left\{\frac{\partial}{\partial\theta}\log\Phi(x|\theta)\right\}^2.$$

But $E'\left\{\frac{\partial}{\partial\theta}\log\Phi(x|\theta)\right\}^2$

$$= \iint\cdots\Phi\times\left(\frac{\partial}{\partial\theta}\log\Phi\right)^2 dx_1\cdots dx_n$$

$$= \iint\cdots\left\{\left(-\frac{\partial^2}{\partial\theta^2}\log\Phi\right)\Phi + \frac{\partial^2\Phi}{\partial\theta^2}\right\}dx_1\cdots dx_n$$

$$= E'\left\{-\frac{\partial^2}{\partial\theta^2}\log\Phi\right\} = -nE\left\{\frac{\partial^2}{\partial\theta^2}\log\varphi\right\}$$

$$= -n\bar{b}$$

$$\therefore \quad E'(\theta-\hat{\theta})^2 = -\frac{1}{n\bar{b}}$$

i.e. $$-\frac{1}{\sigma_{\hat{\theta}}^2} = n\bar{b}$$

provided $E'(\hat{\theta})=\theta$; i.e. provided $\hat{\theta}$ is normally distributed in large samples; the result given by Fisher on p.329.

Hence the well known formula for the variance of the maximum likelihood statistic,

$$-\frac{1}{\sigma_{\hat{\theta}}^2} = \frac{\partial^2}{\partial\theta^2}\log\Phi(x|\hat{\theta}).$$

2·17.

This formula holds only on the assumption that n is large. The assumption is also one of normal distribution of $\hat{\theta}$ and Fisher's proof, which follows, is based on that assumption.

If $\hat{\theta}$ is normally distributed, the probability that, when θ is given, $\hat{\theta}$ should lie in the range $d\hat{\theta}$ can be written in the form

$$dF = \sigma\frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(\hat{\theta}-\theta)^2/\sigma^2}\, d\hat{\theta}$$

whence $$\frac{\partial^2}{\partial\theta^2}\log F = -\frac{1}{\sigma^2}$$

Now F is the total frequency of all samples yielding the value $\hat{\theta}$; and we know that

$$\Phi \propto e^{\frac{1}{2}n\bar{b}(\theta-\hat{\theta})^2},$$

which is constant for all samples yielding $\hat{\theta}$ as the estimate of θ .

Hence as θ varies

$$F \propto e^{\frac{1}{2}n\bar{b}(\theta-\hat{\theta})^2}$$

$$\therefore \quad \frac{\partial^2}{\partial\theta^2}\log F = -\frac{1}{\sigma_{\hat\theta}^2} = n\bar{b},$$

as before.

In large samples, therefore, $\hat\theta$ tends to be normally distributed with variance $\sigma^2 = -1/(n\bar{b})$, where

$$-\bar{b} = E\left\{-\frac{\partial^2}{\partial\theta^2}\log\phi\right\} = E\left\{\frac{\partial}{\partial\theta}\log\phi\right\}^2.$$

2·18·

This quantity $E\left(-\frac{\partial^2}{\partial\theta^2}\log\phi\right) = -\int \phi\cdot\frac{\partial^2}{\partial\theta^2}\log\phi\, dx$,

is what Fisher terms the intrinsic accuracy, i, of the curve as a means of estimating θ. This quantity i may be regarded as the amount of information in a single observation in the sense already explained. If this concept is to be granted then we would expect to find that the amount of information in the sample is ni.

If we have two independent observations from the same population, or from different populations, their joint distribution is given by

$$d\Phi = \phi_1\phi_2\, dx_1\, dx_2.$$

Hence their intrinsic accuracy is

$$-\iint \phi_1\phi_2\frac{\partial^2}{\partial\theta^2}\log(\phi_1\phi_2)\, dx_1\, dx_2$$

$$= -\iint \phi_1\phi_2\left(\frac{\partial^2}{\partial\theta^2}\log\phi_1 + \frac{\partial^2}{\partial\theta^2}\log\phi_2\right) dx_1\, dx_2,$$

which, since $\int \phi_1\, dx_1 = \int \phi_2\, dx_2 = 1$,

reduces to

$$-\int \phi_1 \frac{\partial^2}{\partial\theta^2} \log \phi_1 \, dx_1 - \int \phi_2 \frac{\partial^2}{\partial\theta^2} \log \phi_2 \, dx_2$$

$$= i_1 + i_2 .$$

If the observations are from the same population,

$$\phi_1 = \phi(x_1|\theta) \quad , \quad \phi_2 = \phi(x_2|\theta)$$

and $$i_1 = i_2 = i .$$

Hence the amount of information in a sample of n from the same population equals ni .

2·19.

We have yet to prove that $\hat{\theta}$ satisfies the criterion of efficiency, that is to say we have to show that the variance of $\hat{\theta}$ is not greater than that of any other estimate, T, of θ .

We have

$$\frac{\partial}{\partial\theta} \log \Phi = -nA(\theta - \hat{\theta}),$$

where $-A$ is the limit of $\frac{\partial^2}{\partial\theta^2} \log \Phi$, i.e. of $S\left(\frac{\partial^2}{\partial\theta^2} \log \phi\right)$ taken over the sample.

If T is normally distributed in large samples we have

$$F = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(T-\theta)^2/\sigma^2} dT$$

where F is the frequency of samples yielding T as the estimate of θ .

Let $$F = S(\Phi)$$

Then, since

$$\frac{\partial^2}{\partial\theta^2} \log F = -\frac{1}{\sigma^2} ,$$

we have to make $\frac{\partial^2}{\partial\theta^2}\log F$ a negative quantity as large as possible by suitably grouping the several sorts of samples under the same values of T.

But

$$-\frac{1}{\sigma^2} = \frac{\partial^2}{\partial\theta^2}\log F = \frac{S^2(\Phi')}{S^2(\Phi)} - \frac{S(\Phi'')}{S(\Phi)}$$

And $\frac{S(\Phi')}{S(\Phi)}$ = mean value, within the group, of $-nA(\theta-\hat{\theta})$,

$\frac{S(\Phi'')}{S(\Phi)}$ = " " " " " $-nA+n^2A^2(\theta-\hat{\theta})^2$.

$\therefore \frac{1}{n\sigma^2} = A - nA^2V'(\hat{\theta})$, where $V'(\hat{\theta})$ = variance of $\hat{\theta}$ within the group.

Hence if $T = \hat{\theta}$, then $\hat{\theta}$ is constant within the group and

$$\sigma^2 = \frac{1}{nA}.$$

If T is constant for sets of samples for which $V'(\hat{\theta}) = O(n^{-1})$, then $\frac{1}{n\sigma^2}$ is reduced, and so σ^2 is increased.

Hence the efficiency of $\hat{\theta}$ is proved.

2·20· The validity of the above theorems, and the conditions in which they hold good, were later investigated by Harold Hotelling. The results of his work were published in the Transactions of the American Mathematical Society, vol.32, (1930), and are worth summarising here.

Let θ_0 = true value of θ,

$\hat{\theta}$ = optimum estimate by maximum likelihood,

$\int_{\alpha}^{\beta} \varphi(x|\theta)\, dx$ = probability that an observed x shall lie between α and β,

$\lambda = \log \varphi$.

The following hypotheses can be made:

(a) φ is a continuous function of x, except possibly on a set of values of x of measure zero.

(b) φ is a continuous monotonic function of θ in a θ-interval including θ_0 for all values of x in some interval.

(c) In a θ-interval including θ_0, $\frac{\partial \varphi}{\partial \theta}$ is a continuous function of θ for every value of x. $x^2 \frac{\partial \varphi}{\partial \theta}$ approaches a continuous function of θ as $x \to \pm \infty$. In some x-interval $\frac{\partial \varphi}{\partial \theta} \neq 0$.

The following theorems are proved:

1. If (a) and (b) are satisfied, $\hat{\theta}$ is a consistent statistic. That is to say, there exists a number N such that if $n > N$, then the probability that

$$|\theta - \hat{\theta}| > \delta$$

is less than ϵ. (δ and ϵ are any two positive numbers, and n the number in sample.)

11. If (a) and (c) are satisfied, the distribution of $\hat{\theta}$ approaches normality.

111. If (a) and (c) are satisfied then the variance of $\hat{\theta}$ bears to $E'\left(\frac{\partial}{\partial\theta}\log\Phi\right)^2$ a ratio that approaches unity as n increases.

1V. If (a) and (c) are satisfied for each θ_i in $\phi(x\,|\,\theta_1,\cdots,\theta_n)$, then the joint distribution of the $\hat{\theta}_i$ approaches normality as n increases.

Also, their variances and product variances, v_{ii} and v_{ij}, multiplied by n, approach the elements of a matrix whose inverse is

$$\left[\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log\phi\right)\phi\,dx\right].$$

In other words, after we have satisfied ourselves as to the consistency of $\hat{\theta}$, the application of the formulae for variance, efficiency, etc. depends for its validity upon the satisfaction of conditions (a) and (c).

2·21· Suppose, for example, we take Pearson's Type 111 curve in the form

$$\phi = \frac{1}{a\cdot(p1)}\left(\frac{x-\theta}{a}\right)^p e^{-\frac{x-\theta}{a}},$$

where a and p are given, and we require to estimate θ by the method of maximum likelihood.

The variance of our estimate obtained from

$$-\frac{1}{\sigma_{\hat{\theta}}^2} = E'\left(\frac{\partial^2}{\partial\theta^2}\log\Phi\right)$$

is found to be

$$\frac{a^2(p-1)}{n} .$$

But if p happens to be equal to or less than unity, (as can be the case, since p can take any value greater than -1), then this leads to an absurd result.

The reason for this is simply that the use of the above formula for the variance of $\hat{\theta}$ is justified only when the conditions of theorem lll are satisfied, namely conditions (a) and (c). But condition (c) states that $\frac{\partial\phi}{\partial\theta}$ shall be a continuous function of θ for every x .

When $p = 1$ the curve makes an acute angle at one end with the axis, but as p passes through the value 1 , this angle changes suddenly to a right angle. But θ is the parameter of location, and as θ varies the curve is shifted along the axis. Hence if p has the above critical value then any change in θ will involve a sudden change in the ordinate ϕ . Hence there is, in this case, a discontinuity in $\frac{\partial\phi}{\partial\theta}$, and the use of the formula

$$-\frac{1}{\sigma_{\hat{\theta}}^2} = E'\left(\frac{\partial^2}{\partial\theta^2}\log\Phi\right)$$

is not valid.

2·22· Before proceeding further it will be useful to obtain the

results for more than one parameter. If, for instance, there are two parameters θ_1 and θ_2, and if $\hat{\theta}_1$ and $\hat{\theta}_2$ are their maximum likelihood estimates obtained from

$$\left.\begin{aligned} \frac{\partial}{\partial \theta_1} \log \Phi &= 0, \\ \frac{\partial}{\partial \theta_2} \log \Phi &= 0, \end{aligned}\right\}$$

and if, further, $\hat{\theta}_1$ and $\hat{\theta}_2$ tend to have a normal distribution given by

$$F = \frac{1}{2\pi \sigma_{\hat{\theta}_1} \sigma_{\hat{\theta}_2}} \frac{1}{\sqrt{1-r^2}} \cdot \exp\left\{ \left[\frac{(\hat{\theta}_1-\theta_1)^2}{\sigma_{\hat{\theta}_1}^2} - \frac{2r(\hat{\theta}_1-\theta_1)(\hat{\theta}_2-\theta_2)}{\sigma_{\hat{\theta}_1}\sigma_{\hat{\theta}_2}} + \frac{(\hat{\theta}_2-\theta_2)^2}{\sigma_{\hat{\theta}_2}^2} \right] \frac{1}{1-r^2} \right\} \quad \text{--- (1)}$$

$d\hat{\theta}_1 \, d\hat{\theta}_2$

for given $\hat{\theta}_1$, $\hat{\theta}_2$, where $r \equiv r_{\hat{\theta}_1, \hat{\theta}_2}$

then, as before

$$\frac{\partial^2}{\partial \theta_1^2} \log F = -\frac{1}{\sigma_{\hat{\theta}_1}^2} \cdot \frac{1}{1-r^2}$$

$$\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log F = -\frac{1}{\sigma_{\hat{\theta}_1}\sigma_{\hat{\theta}_2}} \cdot \frac{r}{1-r^2}$$

$$\frac{\partial^2}{\partial \theta_2^2} \log F = -\frac{1}{\sigma_{\hat{\theta}_2}^2} \frac{1}{1-r^2}$$

The Taylor expansion of $\log \Phi$ leads to

$$\log \Phi = C + \frac{1}{2}\left\{ (\theta_1-\hat{\theta}_1)^2 \, n E\left(\frac{\partial^2 L}{\partial \theta_1^2}\right) + 2(\theta_1-\hat{\theta}_1)(\theta_2-\hat{\theta}_2) \, n E\left(\frac{\partial^2 L}{\partial \theta_1 \partial \theta_2}\right) + (\theta_2-\hat{\theta}_2)^2 \, n E\left(\frac{\partial^2 L}{\partial \theta_2^2}\right) \right\},$$

--- (2)

where $n E\left(\frac{\partial^2 L}{\partial \theta_1^2}\right) \equiv n E\left\{\frac{\partial^2}{\partial \theta_1^2} \log \varphi(x|\theta_1,\theta_2)\right\}$, etc.

$\therefore \log F = C' + \frac{1}{2}\{\cdots\cdots\cdots\cdots\}$, as before.

Hence $-\frac{1}{\sigma_{\hat{\theta}_1}^2} \cdot \frac{1}{1-r^2} = \overline{\frac{\partial^2}{\partial \theta_1^2} \log \Phi}$, in the limit;

$$-\frac{1}{\sigma_{\hat{\theta}_1}\sigma_{\hat{\theta}_2}} \cdot \frac{r}{1-r^2} = \overline{\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log \Phi} ,$$

$$-\frac{1}{\sigma_{\hat{\theta}_2}^2} \cdot \frac{1}{1-r^2} = \overline{\frac{\partial^2}{\partial \theta_2^2} \log \Phi} .$$

But the two forms in which $\log F$ can be written (i.e. (1) and (2)) lead to the following rule for obtaining the variances:

The variances and product variances σ_{ii} and σ_{ij}, each multiplied by n, approach in the limit the elements of a matrix whose inverse is

$$\left[\int_{-\infty}^{\infty} \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log \varphi \right) \cdot \varphi \cdot dx \right]$$

In practice $\sigma_{\hat{\theta}}^2$ is found by dividing the Hessian determinant of $\log \Phi$, with respect to the parameters, into the corresponding minor.

2·23· All the preceding sections have been developed on the

assumption that n is large; and we have shown that in such a case the maximum likelihood estimate uses an amount of information which tends to ni, where i is the amount of information in a single observation. In other words, in the limit the maximum likelihood statistic uses the whole of the information in the sample. But, it happens that there is a fairly wide class of statistics which use the whole of the relevant information when n is finite. To such statistics Fisher has applied the term "sufficient." To complete a theory of estimation it is necessary

(1) To find the conditions for the existence of sufficient statistics;

(2) To show that any proposed method yields such statistics when they exist;

(3) To find an expression for the <u>loss</u> of information, through the use of a non-sufficient statistic when no sufficient statistic exists;

(4) To find a process whereby some or all of this "lost" information can be recovered.

2·24· A more precise meaning of the term "sufficient" may be given as follows:

If T_1 is a sufficient statistic, and T_2 any other statistic,

then the joint distribution of T_1 and T_2 must be such that, when T_1 is given, the distribution of T_2 does not involve θ. In this case, knowledge of T_2 can add nothing to our knowledge of θ, and T_1 must contain all the relevant information concerning θ. In other words if $F(T_1, T_2 | \theta)$ is the distribution of T_1, T_2, then

$$F(T_1, T_2 | \theta) = f_1(T_1 | \theta) f_2(T_2 | T_1).$$

As an example we have the mean, $\bar{x}$, of the distribution, $e^{-nm} \frac{(nm)^{n\bar{x}}}{(n\bar{x})!}$.

The probability, $p(x_1, \cdots, x_n)$ of drawing in order a particular sample $x_1, \cdots, x_n$ is

$$e^{-nm} \frac{m^{n\bar{x}}}{x_1! x_2! \cdots x_n!} = e^{-nm} \frac{(nm)^{n\bar{x}}}{(n\bar{x})!} \frac{(n\bar{x})!}{n^{n\bar{x}} x_1! \cdots x_n!},$$

where $e^{-nm} \frac{(nm)^{n\bar{x}}}{(n\bar{x})!}$ = probability of scoring total $n\bar{x}$,

$$= p(nT | m), \text{ say;}$$

and $\frac{(n\bar{x})!}{n^{n\bar{x}} \cdot x_1! \cdots x_n!}$ = probability that, given the total $n\bar{x}$, i.e. given nT, the order should be $x_1, \cdots, x_n$,

$$= p(x_1, \cdots, x_n | nT), \text{ say.}$$

Hence another way of expressing is

$$p(x'|\theta) = p(T|\theta) \times p(x'|T),$$

where $x' \equiv$ vector $(x_1, \ldots, x_n)$.

From this it is obvious that, when a sufficient statistic exists, the method of maximum liklihood will give it; for

$$\frac{\partial}{\partial\theta} \log p(x'|\theta) = \frac{\partial}{\partial\theta} \log p(T|\theta).$$

Later we will find a general expression for the form of all distributions which can yield sufficient statistics for some or all of its unknown parameters.

2·25 The existence of such sufficient statistics, which utilise the whole of the information available in the sample when n is finite, suggests a new definition of "efficiency" which will apply to finite samples without the assumption of normal distribution of the statistics. The definition is as follows:

"The efficiency of a statistic is the ratio of the intrinsic accuracy of its random sampling distribution to the amount of information in the data from which it has been derived."

The intrinsic accuracy = mean value of

$$\left(\frac{1}{\phi}\,\frac{\partial\phi}{\partial\theta}\right)^2.$$

If T is to be a sufficient statistic; i.e. if there is to be no loss of information through using T, it is obvious that,

as T must be a maximum likelihood statistic, all samples yielding T as an estimate must have a constant value for $\frac{\partial}{\partial\theta}\log\Phi$ when θ is given; and no matter what the value of θ.

If, on the other hand, $\frac{\partial}{\partial\theta}\log\Phi$ is constant in a set of samples for a particular value of θ, but is no longer constant within the set for other values of θ, then no sufficient statistic exists, and a loss of information will result from the use of a single estimate.

2·26 To find an expression for the loss of information through using a non-sufficient statistic, we proceed as follows:

If in large samples we get observed values $x_1, \cdots, x_n$ where the expected values are $m_1, \cdots, m_n$, then we have

$$\log\Phi = S(x \log m),$$

$$\frac{\partial}{\partial\theta}\log\Phi = S\left(x\,\frac{m'}{m}\right),$$

$$\frac{\partial^2}{\partial\theta^2}\log\Phi = S\left\{x\left(\frac{m''}{m} - \frac{m'^2}{m^2}\right)\right\};$$

and also, as we have seen, we have, for n large

$$\frac{\partial}{\partial\theta}\log\Phi = (\theta - \hat{\theta})\frac{\partial^2}{\partial\theta^2}\log\Phi$$

Hence the variance of $\frac{\partial}{\partial\theta}\log\Phi$ in a set of samples for which $\hat{\theta}$ is constant is equal to $(\theta-\hat{\theta})^2$ times the variance of $\frac{\partial^2}{\partial\theta^2}\log\Phi$. Hence the total loss of information is

given by $V(\hat{\theta})$ times the general variance within such sets.

From this Fisher deduces the expression for the loss of information in the form

$$\frac{S\left\{\frac{1}{m}\left(m'' - \frac{m'^2}{m}\right)^2\right\}}{S\left(\frac{m'^2}{m}\right)} - \frac{1}{n} S\left(\frac{m'^2}{m}\right) - \frac{S^2\left\{\frac{m'}{m}\left(m'' - \frac{m'^2}{m}\right)\right\}}{S^2\left(\frac{m'^2}{m}\right)}.$$

2·24. The application of this formula to, say, the Cauchy distribution

$$d\phi = \frac{1}{\pi} \frac{dx}{1 + (x-\theta)^2}$$

yields the result that, for large n, the maximum likelihood estimate given by

$$S\left\{\frac{x_r - \hat{\theta}}{1 + (x_r - \hat{\theta})^2}\right\} = 0$$

rejects an amount of information equivalent to that in $2\frac{1}{2}$ observations.

The mean, on the other hand, as we have seen earlier, can be shown to have exactly the same sampling distribution as the distribution of a single observation. In other words, the mean uses only as much information as it contained in a single observation. Hence the mean is quite useless as a statistic for estimating θ.

The median, on the other hand, has a sampling variance of

$\frac{\pi^2}{4n}$, whilst the variance of $\hat{\theta}$ is $\frac{2}{n}$. Hence the efficiency of the median is $\frac{2}{n} / \frac{\pi^2}{4n} = 8/\pi^2$, or approximately 81%. Hence the median is much superior to the mean as a statistic for locating the curve.

2·28. The variance of a non-sufficient statistic, T, is compounded from two causes:

(1) The sampling error $T - \theta$ which, provided T is consistent, tends to the same limit when n is large, no matter how T is derived.

(2) The deviation of T from $\hat{\theta}$ the maximum liklihood (minimum variance) statistic.

The method of minimum χ^2 developed by Kirstine Smith (*q.v.*) is criticised by Fisher from this standpoint.

For $$\chi^2 = S\left\{\frac{(x-m)^2}{m}\right\}.$$

Hence for minimum χ^2 we have

$$S\left\{\frac{x^2-m^2}{m^2} \cdot \frac{\partial m}{\partial \theta}\right\} = 0;$$

whilst for maximum liklihood

$$S\left\{\frac{x-m}{m} \cdot \frac{\partial m}{\partial \theta}\right\} = 0.$$

Since $$\frac{x^2-m^2}{m^2} \Big/ \frac{x-m}{m} = \frac{x+m}{m},$$

and $\frac{x+m}{m} \rightsquigarrow 2$ when n is large

minimum χ^2 yeilds an efficient statistic.

But since $x^2 - m^2 = (x-m)^2 + 2m(x-m)$,

the deviation from $\frac{\partial}{\partial \theta} \log \Phi$ will be

$$\frac{1}{2} S \left\{ \frac{(x-m)^2}{m^2} \right\} \frac{\partial m}{\partial \theta}$$

and the variance of this can be shown to be

$$\frac{1}{2} S\left(\frac{m'^2}{m^2}\right) - \frac{1}{2} \frac{S^2\left(\frac{m'^3}{m^2}\right)}{S^2\left(\frac{m'^2}{m}\right)} .$$

This quantity remains finite as n increases but tends to infinity as the number of classes is increased. Hence the method breaks down for fine grouping. Actually this is to be expected, since χ^2 is an approximation, only valid when the number of observations in each class is large.

2·29 When the use of a non-sufficient statistic leaves a measureable amount of the information unused, this information may be (in theory at least) partly or wholly recovered by calculating what Fisher calls "ancillary statistics" which tell us how good an estimate we have made of the parameter. "Ancillary statistics," says Fisher "are only useful when different samples of the same size can supply different amounts of information, and serve to distinguish those which supply more from those which supply less." Such additional information

can be obtained from other characteristics of the likelihood function such as its second and higher derivation at the maximum.

Sometimes the ancillary information may be obtained from the configuration of the sample, (e.g. Fisher's use of marginal sums in the 4-way table, ["Two New Properties of Mathematical Liklihood," Proc. Roy. Soc., A144, (1934), pp48-51.)] But in general it appears that the whole of the information can be recovered only by the use of the whole course of the likelihood function. In effect this means dropping the whole notion of estimation, or the reduction of data, and reverting to the sample itself.

2·30

If we have a curve of the form

$$d\phi = \text{const.} \times e^{f(\xi)}\, d\xi \quad , \quad \xi = \frac{x-\theta_1}{\theta_2} ,$$

the estimation of θ_1 and θ_2 is called the location and scaling of the frequency curve; θ_1 being the parameter of location and θ_2 the parameter of scaling.

We have

$$\frac{\partial}{\partial \theta_1} \log \Phi = -\frac{1}{\theta_2} S(f') \quad , \quad f' = \frac{\partial f}{\partial \xi} , \qquad \ldots (1)$$

$$\frac{\partial}{\partial \theta_2} \log \Phi = -\frac{1}{\theta_2} S(\xi f') , \qquad \ldots (2)$$

$$\frac{\partial^2}{\partial \theta_1^2} \log \Phi = \frac{1}{\theta_2^2} S(f''), \quad \cdots (3)$$

$$\frac{\partial^2}{\partial \theta_2^2} \log \Phi = \frac{1}{\theta_2^2} S(\xi^2 f'' - 1), \quad \cdots (4)$$

$$\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log \Phi = \frac{1}{\theta_2^2} S(\xi f''). \quad \cdots (5)$$

(3) gives $-\frac{1}{\sigma_{\hat{\theta}_1}^2} = \frac{n E(f'')}{\theta_2^2}$ (e.g. normal curve)

(4) on the other hand does not give

$$-\frac{1}{\sigma_{\hat{\theta}_2}^2} = \frac{n}{\theta_2^2} E(\xi^2 f'' - 1),$$

unless $\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log \Phi = 0$, in which case the Hessian determinant becomes

$$\begin{vmatrix} \frac{\partial^2}{\partial \theta_1^2} \log \Phi & 0 \\ 0 & \frac{\partial^2}{\partial \theta_2^2} \log \Phi \end{vmatrix}, \quad \text{when } n \text{ is large.}$$

But $\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log \Phi$ can be made zero by subtracting from ξ the quantity

$$\frac{E(\xi f'')}{E(f'')}.$$

For example in the type III distribution

$$\phi = \frac{1}{\theta_2 \, p!} \left(\frac{x-\theta_1}{\theta_2} \right)^p e^{-\frac{x-\theta_1}{\theta_2}},$$

where p is given,

$$f = p \log \xi - \xi .$$

In this case we find

$$E(\xi f'') = 1 \quad , \quad E(f'') = -\frac{1}{p-1}$$

$$\frac{E(\xi f'')}{E(f'')} = -(p-1) .$$

Hence we write

$$f = p \log (\xi + p-1) - (\xi + p-1)$$

and find

$$\sigma^2_{\hat{\theta}_2} = \frac{a^2}{2n} .$$

2·31· So far we have been dealing with continuous distributions; but the method of maximum liklihood applies equally well to discontinuous distributions, as the following will show:

Let p_s = probability of an observation falling into cell s ,

and n_s = number of observations actually falling into that cell in a sample of N .

Then

$$S(\log \phi) = S(n_s \log p_s).$$

Let

$$L = S(n_s \log \frac{n_s}{\bar{n}_s}),$$

where $\bar{n}_s = p_s N.$

Then L differs only be a constant from $\log \Phi$ with sign reversed.

Hence the problem of finding the maximum of $\log \Phi$ is the same as the problem of finding the minimum of L.

Hence $$\frac{\partial L}{\partial \theta} = -S\left(\frac{n_s}{\bar{n}_s} \frac{\partial \bar{n}_s}{\partial \theta}\right) = 0.$$

For example, in the Poisson distribution,

$$e^{-m}\left(1, m, \frac{m^2}{2!}, \ldots \frac{m^x}{x!}, \ldots\right)$$

$$S\left\{\frac{\partial}{\partial m}(-m + x \log m)\right\} = 0,$$

$$\therefore \hat{m} = \bar{x}.$$

Also

$$-\frac{1}{\sigma^2_{\hat{m}}} = S\left(-\frac{x}{m^2}\right) = -\frac{n}{m}$$

$$\therefore \sigma^2_{\hat{m}} = \frac{m}{n}.$$

2·32

Although the finding of maximum likelihood estimates is simple

in theory, in practice they are not so easy to use. We have already seen this in the case of the Cauchy distribution where the maximum liklihood estimate of the unknown parameter depended upon the approximate solution of an equation of degree $2n+1$.

The same difficulty arises in the application of maximum likelihood to curve fitting, especially in the case where no sufficient statistic exists.

A worked example of the application of the method to the problem of curve fitting was given by R. S. Koshal, [J.R.S.S. (1933), Part 11, P303 ,]and was the occasion of controversy between Professor Fisher and the late Professor Karl Pearson. It is perhaps rather unfortunate that on one side this controversy turned in part on the accuracy or rather the lack of accuracy in the calculations not only of Koshal but of Pearson himself in his efforts to correct Koshal. The controversy in itself was important and interesting enough and a brief account of it will be given shortly. However, the unfortunate part of it was that the underlying theory itself was lost sight of, and a restatement of the theory is necessary here.

Suppose that we are concerned with fitting data to a curve possessing two parameters θ_1 and θ_2. Suppose that T_1 and T_2 are insufficient estimates of θ_1 and θ_2 obtained, say, by the method of moments, and that T_1' and T_2' are the maximum

likelihood estimates to a first approximation. We require to find the corrections $(T_1' - T_1)$ and $(T_2' - T_2)$ which must be added to T_1 and T_2 to give the maximum likelihood estimates.

If $L \equiv \log \Phi$, we have

$$(T_1' - T_1)\frac{\partial^2 L}{\partial \theta_1^2} + (T_2' - T_2)\frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} = -A_1 ,$$

$$(T_1' - T_1)\frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} + (T_2' - T_2)\frac{\partial^2 L}{\partial \theta_2^2} = -A_2 ,$$

where A_1 and A_2 are the discrepancies from zero of $\frac{\partial L}{\partial \theta_1}$ and $\frac{\partial L}{\partial \theta_2}$, whose numerical values are calculated by putting T_1 and T_2 for θ_1 and θ_2 in $\frac{\partial L}{\partial \theta_1}$ and $\frac{\partial L}{\partial \theta_2}$; Also, in evaluating $\frac{\partial^2 L}{\partial \theta_1^2}$ etc., we use T_1 for θ_1 and so on.

The above equations are derived, in the manner used in previous sections, from the Taylor expansion for $\frac{\partial L}{\partial \theta}$ taken to a first approximation, and serve to determine the required corrections provided we can find the numerical values of the differential coefficients.

2·33. The difficulties arise from the fact that, whereas we can always find the numerical value of L from $L = S(n_s \log p_s)$ where n_s and p_s are respectively the number observed in class S , and the probability of an observation falling within that class, we cannot always calculate the values of $\frac{\partial^2 L}{\partial \theta_1^2}$,

and so on. In such a case we proceed as follows:

Let L_{00} = value of the liklihood calculated from the moment solution.

Then keep θ_2 constant and give θ_1 two small increments in turn so as to obtain two values L_{10} , L_{20} on either side of the true maximum. Similarly, keeping θ_1 constant, obtain L_{01} and L_{02} . Finally giving one increment each to θ_1 and θ_2 obtain L_{11} .

If now our desired corrections, in terms of the chosen increments as units, are x and y , then the value of L obtained by using ξ and η increments respectively must be a maximum when $\xi = x$ and $\eta = y$. That is to say the likelihood curve, near its maximum may be taken to be a parabola, and may be expressed in the form

$$L = c - a(x-\xi)^2 - 2h(x-\xi)(y-\eta) - b(y-\eta)^2,$$

where ξ and η take the values 0, 1, 2 in the trial cases evaluated; i.e. for L_{00} , $\xi = 0$, $\eta = 0$,

for

$$L_{10} \quad , \quad \xi = 1 \quad , \quad \eta = 0 ,$$

$$L_{20} \quad , \quad \xi = 2 \quad , \quad \eta = 0 ,$$

$$L_{01} \quad , \quad \xi = 0 \quad , \quad \eta = 1 ,$$

$$L_{02} \quad , \quad \xi = 0 \quad , \quad \eta = 2 ,$$

$$L_{11} \quad , \quad \xi = 1 \quad , \quad \eta = 1 .$$

Substituting these in the expression for L we get

$$\left.\begin{aligned} L_{10} - L_{00} &= a(2x-1) + 2hy \\ L_{01} - L_{00} &= 2hx + b(2y-1) \end{aligned}\right\} \text{ for } x \neq y.$$

For a, b, h we have

$$\left.\begin{aligned} L_{20} - 2L_{10} + L_{00} &= -2a, \\ L_{02} - 2L_{01} + L_{00} &= -2b, \\ L_{11} + L_{00} - L_{01} - L_{10} &= -2h. \end{aligned}\right\}$$

In a word the method consists in finding the increments which must be added to our original estimates in order to make the liklihood a maximum, on the assumption that near the maximum the liklihood curve assumes the form of a parabola.

2·34. Koshal applied this method to fitting a type I curve to some given data, and thought he proved that the curve obtained was a better fit by the χ^2 criterion than that obtained by the method of moments. In practice the method is a very tedious one, involving, as it does, first of all the calculation of θ_1 and θ_2 by the method of moments, and then the calculation of the various values of the liklihood.

In criticism of this, Karl Pearson, [Biom., 28, (1936),] maintained that the value of the example was negated by three serious blunders - (1) Raw (uncorrected) moments were used; (2) two of these were wrongly calculated; (3) the range of the curve was confused with the abscissa of the end-point. Correcting these mistakes, and then, comparing the results, Pearson showed by the criteria of both χ^2 and [(likelihood) the method of moments was superior.

Further he stated that, if the χ^2 test was accepted as a test of goodness of fit, then it was incumbent on Fisher to prove that the method of maximum liklihood should make χ^2 a minimum - and, he averred, that Fisher was unable to do this.

The dispute seems then to reduce to the comparison of minimum χ^2 and maximum likelihood as criteria of goodness of fit. Actually the likelihood is no more a criterion of goodness of fit than is χ^2 an efficient method of estimation. Pearson claimed that the χ^2 test, based as it is upon actual, concrete comparison between theoretical and observed values, is much more readily to be accepted by the practical statistician than the method of maximum liklihood, which is based upon a somewhat vague and arbitrary conception of "best values." All of which seems to be very wide of the mark and very much beside the point.

In "Biometrika", Vol.XI, (1915), Dr. Kirstine Smith gives a method for obtaining the "best" value of the standard deviation, σ , of a series of observations from a normal population. This method is based upon the simple expedient of finding that value of σ which makes χ^2 a minimum - and she finds the values σ = 2·355,860, χ^2 = 9·720. Fisher, in his 1922 paper, using maximum likelihood finds σ = 2·26437 which is very close to that obtained from the usual method with Shepherd's correction, σ = 2·26421.

In criticism of this Pearson points out (1) that Fisher should have cast out the last observation and (2) that Fisher's method raised χ^2 from 9·720 to 11·828.

In regard to the first point it is rather incomprehensible why Fisher should be asked to cast out the extreme observation whilst K. Smith is allowed to retain it. In regard to the second point, the fact that Fisher's method yields χ^2 = 11·828, while K. Smith's method yields χ^2 = 9·720, proves nothing in itself. K. Smith has departed from the conventional formula for σ in order to make χ^2 a minimum. We are still, apparently, as much in the dark as ever as to the superiority of one method over the other. All that Pearson can claim for K. Smith's method is that the process for minimising χ^2 yields a "curve which gives us as nearly as possible, by the χ^2 test, the grouped

frequencies observed". In other words, the process of minimising χ^2 yields the minimum χ^2 ! Logically this is as helpful as the conclusion that, since water boils at the temperature of 100 degrees centigrade, we can tell whether water is at the temperature of 100 degrees by observing whether it not it is boiling.

Besides this, χ^2 is itself an approximation applicable only when (a) the observed and theoretical frequencies are large, and (b) the difference between them is of a lower order of magnitude. Neither of these two conditions is satisfied in that very class (the extreme) which Pearson said Fisher should have rejected ! If this last class has vitiated Fisher's result, how much more so has it vitiated K. Smith's result ?

Actually, however, it happens that for large samples and when the conditions (a) and (b) above are satisfied there is close agreement between the results obtained by maximising the likelihood and those obtained by minimising χ^2.

To return for a moment to Koshal's example. Pearson claimed that, having corrected the three mistakes made by Koshal, he obtained results showing that the method of moments gave a value of χ^2 that was less than that given by the corrected estimates. Unfortunately it happens that Pearson himself committed three new blunders in the process of

refuting Koshal. It would be tedious to pursue these matters further. The mistakes made by Pearson are fully elaborated by R. A. Fisher, [Annals of Eugenics, Vol.Vll, Part lV, (1937), pp.303-308.] Suffice to say that Fisher did a lot of damage to Pearson's claim to be the constant champion of the "practical statistician."

2·35· To sum up the controversy let us say first of all that, scientifically speaking, it is a very risky procedure to elaborate a method of estimation (such as the χ^2 method) that is based upon a certain principle, and then to test our results by an application of the selfsame principle. In all scientific work it is generally admitted that the most satisfactory test of any results is a test that is independent, as far as possible, of the method by which those results have been obtained. Or, on the other hand, if the satisfaction of a certain condition is judged necessary on the basis of practical experience, or on other grounds, let us incorporate that condition into our method. Such a condition might be say, a minimum variance condition, but not a minimum χ^2 condition for the reasons already given. If we incorporate into our method of estimation a minimum variance condition, we will not conclude that our estimate is the "best" because it has minimum variance, but that a minimum variance estimate will be the best in the sense of being the most reliable, a concept which is not peculiar to the theory of estimation alone.

CHAPTER THREE

THE METHOD OF CONSISTENT MINIMUM VARIANCE

3.01

Let us now return to our original statement of the problem of estimation.

We have seen that if $T(x_1, \cdots, x_n)$ is to be accepted as the best estimate of an unknown parameter θ specifying a population of given form, then

(1) T must be a consistent statistic in the sense that there exists a number N such that, when $n > N$, the probability that

$$|\theta - T| > \epsilon$$

is less than δ;

(2) T must be such that its variance is not greater than that of any other estimate, T_1, of θ.

3.02 Unbiassed estimates

Instead of making a third assumption, such as that of maximum likelihood, and showing that the solution satisfies (1) and (2) above, we will state a new and more rigorous consistency condition, and combine it with condition (2) to obtain a unique solution.

$T(x_1, \cdots, x_n)$ being a symmetric function of the observations $x_1, \cdots, x_n$, let our conditions be:

(a) T is an "unbiassed estimate" of θ, in the sense that

θ is equal to the expectation of T over all samples of n observations; i.e.

$$\iiint \cdots T(x_1, \cdots, x_n) \prod_{i=1}^{n} \phi(x_i)\, dx_i = \theta$$

where $\phi(x)$ is the elementary probability law for the x_i.

As an example of a biassed estimate we have

$$\sigma_1^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} (x_i),$$

as an estimate of σ^2 in the normal curve

$$\phi(x)\, dx = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-m)^2/\sigma^2}$$

For, as we have seen, the expectation of σ_1^2,

$$E(\sigma_1^2) = (1 - \tfrac{1}{n})\sigma^2;$$

and so, in finite samples, σ_1^2 shows a negative bias of $\frac{\sigma^2}{n}$. When n becomes infinite, this bias becomes zero, and accordingly the usual consistency condition is satisfied.

3·03 Postulate of Minimum Variance

(b) The sampling variance of T in samples of n is to be a minimum, subject to condition (a); i,e,

$$\iiint \cdots (T-\theta)^2 \prod_{i=1}^{n} \phi(x_i)\, dx_i = \min.$$

3·04 Locating the Normal Curve

For example, let us estimate m in

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-m)^2}.$$

By (a)

$$\int_{-\infty}^{\infty}\int\cdots\int -(T-m)\prod\phi(x)\,dx = 0$$

By (b)

$$\int_{-\infty}^{\infty}\int\cdots\int (T-m)^2\prod\phi(x)\,dx = \text{min.},$$

where

$$\prod\phi(x) = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\Sigma(x_i-m)^2}.$$

Regard m as an arbitrary parameter, and differentiate under the integral:

$$\int_{-\infty}^{\infty}\int\cdots\int \left\{-2(T-m)-(T-m)^2(\Sigma x - nm)\right\}\prod\phi(x)\,dx = 0;$$

i.e.

$$\int_{-\infty}^{\infty}\int\cdots\int (T-m)^2(\Sigma x - nm)\,\pi\phi(x)\,dx = 0, \text{ for all } m,$$

since

$$\int_{-\infty}^{\infty}\int\cdots\int (T-m)\,\pi\phi(x)\,dx = 0.$$

Hence, if $z = x-m$, we have

$$\int_{-\infty}^{\infty}\int\cdots\int \left\{T(z+m)-m\right\}^2(\Sigma z)\prod\phi_1(z)\,dz = 0,$$

for all m.

Hence $\left\{T(z+m)-m\right\}$ must be independent of m, and so must be linear. Being a symmetric function of n variables, T must therefore be

$$\frac{1}{n}\Sigma(z+m),$$

i.e.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i),$$

the mean of sample.

3·05. The General Solution

To consider the general solution.

Let

$$\Phi(x'|\theta) = \phi(x_1|\theta)\,\phi(x_2|\theta)\cdots\cdots\phi(x_n|\theta)$$

where

$$x' \equiv \text{vector } (x_1, x_2, \cdots\cdots, x_n).$$

We have

$$\int\int\cdots\int \Phi(x'|\theta)\,dx' = 1 \quad ,$$

$$\int\int\cdots\int (T-\theta)\,\Phi(x'|\theta)\,dx' = 0 \quad ,$$

$$\int\int\cdots\int (T-\theta)^2\,\Phi(x'|\theta)\,dx' = \min.$$

The solution of this is most easily found by the Calculus of Variations. The "Euler Equation" gives

$$2(T-\theta)\Phi - \lambda\Phi_\theta = 0 \quad , \quad \left\{\Phi_\theta = \frac{\partial}{\partial\theta}\Phi(x'|\theta)\right\},$$

i.e.

$$T = \theta - \tfrac{\lambda}{2}\Phi_\theta/\Phi \quad ,$$

where

$$\lambda \equiv \lambda(\theta)$$

and $\lambda(\theta)$ is independent of x', but dependent on θ.
The problem is "isoperimetrical," of positive definite type.

$$T = \theta - \tfrac{\lambda}{2}\Phi_\theta/\Phi$$

is independent of θ. Hence

$$\tfrac{\lambda}{2}\Phi_\theta/\Phi = \theta - T(x').$$

A solution is possible provided

$$\Phi_\theta / \Phi = \frac{2}{\lambda(\theta)} \{ \theta - T(x_1, \cdots, x_n) \},$$

and in that case is $T(x_1, \cdots, x_n)$.

When there are several parameters, and

$$\Phi = \Phi(x_1, \cdots, x_n | \theta_1, \cdots, \theta_K)$$

there are K conditions

$$\frac{\partial \Phi}{\partial \theta_r} / \Phi = \frac{2}{\lambda_r(\theta_1, \cdots, \theta_K)} \{ \theta_r - T_r(x_1, \cdots, x_n) \},$$

$$r = 1, 2, \cdots, K;$$

which must be simultaneously satisfied.

3·06. Examples

I To estimate m in $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-m)^2}$.

Here

$$\Phi = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\sum_{r=1}^{n}(x_r - m)^2}$$

$$\Phi_m / \Phi = \frac{\partial}{\partial m} \log \Phi = \Sigma x - nm,$$

$$= -\frac{2}{\lambda(m)} (T - m) ;$$

$$\therefore \quad T = m - \frac{\lambda(m)}{2} \{ \Sigma x - nm \}$$

and this is to be independent of m. Thus

$$\frac{\lambda(m)}{2} = -\frac{1}{n}$$

giving

$$T = \frac{1}{n} \Sigma x = \overline{x}$$

II To estimate m in the continuous Poisson curve,

$$\phi = e^{-m}\frac{m^x}{\Gamma(x+1)} .$$

$$\Phi = e^{-nm}\frac{m^{\Sigma(x)}}{\prod_r\{\Gamma(x_r+1)} ;$$

$$\frac{\partial}{\partial m}\log\Phi = -n + \Sigma(x)/m ,$$

$$= \frac{2}{\lambda(m)}\left\{m - T\right\} .$$

$$\therefore \frac{\Delta(m)}{2} = -\frac{m}{n} ,$$

giving

$$T = \frac{\Sigma(x)}{n} ,$$

the mean of observations.

III To estimate σ^2 in

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\, e^{-\frac{1}{2}(x-m)^2/\sigma^2} .$$

The difficulty involved here arises from the fact that we are really concerned with the simultaneous estimation of m and σ. If we proceed at first without any assumptions as to the dependence of σ on m, we get

$$\Phi = (2\pi)^{-\frac{1}{2}n}(\sigma^2)^{-\frac{1}{2}n}\, e^{-\frac{1}{2}\Sigma(x-m)^2/\sigma^2} ;$$

$$\therefore \frac{\partial}{\partial\sigma^2}\log\Phi = -\frac{n}{2\sigma^2} + \frac{1}{2}\frac{\Sigma(x-m)^2}{\sigma^4} ,$$

$$= \frac{2}{\lambda(\sigma^2)}\left\{\sigma^2 - T\right\} .$$

$$\therefore T = \sigma^2 + \frac{\Delta(\sigma^2)}{2}\left\{\frac{n}{2\sigma^2} - \frac{1}{2}\frac{\Sigma(x-m)^2}{\sigma^4}\right\} ;$$

$$\therefore \quad \frac{\lambda(\sigma^2)}{2} = -\frac{2\sigma^4}{n}$$

would give

$$T = \frac{1}{n}\sum_{r=1}^{n}(x_r - m)^2.$$

But this is useless, due to the fact that m is unknown. (Cf. the corresponding situation in "maximum likelihood.")

If, on the other hand, we assume that m is known, and has been estimated from $\frac{1}{n}\Sigma(x)$, then we know that the expectation of T is equal to $\frac{n-1}{n}$ times σ^2. In other words, in estimating m we have lost one degree of freedom. Hence we desire

$$T(x_1, \cdots, x_n)$$

such that

$$\int\int\cdots T \cdot \Phi(x_1, \cdots, x_{n-1})\, dx_1 \cdots dx_{n-1} = \sigma^2,$$

where

$$\Phi = (\sigma^2)^{-\frac{1}{2}(n-1)}(2\pi)^{-\frac{1}{2}(n-1)} e^{-\frac{1}{2}\Sigma(x_r - \bar{x})^2/\sigma^2}$$

Hence

$$-\frac{n-1}{2\sigma^2} + \frac{1}{2}\frac{\Sigma(x_r - \bar{x})}{\sigma^4} = \frac{2}{\lambda(\sigma^2)}\left[\sigma^2 - T(x_1 \cdots x_n)\right];$$

and

$$\frac{\lambda}{2} = -\frac{2\sigma^4}{n-1}, \qquad T = \frac{1}{n-1}\sum_{r=1}^{n}(x_r - \bar{x})^2.$$

It is worth noting here that the method of maximum likelihood leads to the result

$$\hat{\sigma}^2 = \frac{1}{n}\Sigma(x_r - \bar{x})^2,$$

and it would seem that an important point in the theory of simultaneous estimation is slightly clouded over by that method.

This point will be discussed in a later section.

IV Let us try to estimate, not for σ^2, but for σ in

$$\phi = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-m)^2/\sigma^2}.$$

Without any considerations as to bias, let us try to solve

$$\iiint \cdots T \cdot \Phi \cdot dx_1 \cdots dx_n = \sigma,$$

$$\iiint \cdots (T-\sigma)^2 \cdot \Phi \cdot dx_1 \cdots dx_n = \min.$$

$$\Phi = (2\pi)^{-\frac{1}{2}n} \sigma^{-n} e^{-\frac{1}{2}\Sigma(x_r-m)^2/\sigma^2},$$

$$\frac{\partial}{\partial\sigma} \log \Phi = -\frac{n}{\sigma} + \frac{\Sigma(x_r-m)^2}{\sigma^3}$$

$$= \frac{2}{\lambda(\sigma)} \{\sigma - T\};$$

$$\therefore \quad T = \sigma + \frac{\lambda(\sigma)}{2} \left\{ \frac{n}{\sigma} - \frac{\Sigma(x_r-m)^2}{\sigma^3} \right\}.$$

It is quite obvious that, even if m is regarded as known, we cannot find a function $\frac{\lambda(\sigma)}{2}$ which will make T independent of σ. Hence we cannot estimate for σ, but we must estimate for σ^2, and put $T = \sqrt{T'}$, where T' is our estimate of σ^2.

V The "Cauchy" distribution,

$$\phi = \frac{1}{\pi} \frac{1}{1+(x-m)^2}$$

To estimate for m :

$$\log \Phi = -n \log \pi - \sum_{r=1}^{n} \log \{1+(x_r-m)^2\},$$

$$\frac{\partial}{\partial m}\log\Phi = \sum \frac{2(x_r-m)}{1+(x_r-m)^2}$$

$$= \frac{2}{\lambda(m)}\{m-T\}$$

$$\therefore \quad T = m - \frac{\lambda(m)}{2}\left\{\sum \frac{2(x_r-m)}{1+(x_r-m)^2}\right.$$

and this is to be independent of m.

It is obvious that no solution exists.

3·07 Comparison with Maximum Likelihood

When $\frac{\partial}{\partial\theta}\log\Phi$ separates into the form

$$\frac{2}{\lambda(\theta)}\{F(\theta) - f(x_1,\cdots,x_n),$$

then $F(\theta)$ is the function which will have to be used for estimating θ; and the estimate given by

$$F(\theta) = f(x_1,\cdots,x_n)$$

will be precisely that given by Fisher's method of maximum likelihood, which puts

$$\frac{\partial}{\partial\theta}\log\Phi = 0.$$

As we have seen, Fisher has proved that (i) where the number in sample tends to infinity, and then (ii) the sampling distribution is normal, the maximum likelihood solution has less variance of error than any other solution. The standpoint we have adopted shows that, in certain cases at least, e.g. examples I, II and III, these assumptions are unnecessary.

3·08 "Basic" Parametric Functions

Two further points of importance are brought out by examples

IV and V .

By example IV we see that we cannot estimate directly for σ in the normal curve, but must estimate for σ^2. In fact it seems certain that there is only one function of the parameter θ for which we can estimate, except of course in the trivial case where, if Θ is that function, we can estimate for Θ itself or for $a\Theta + b$, where a and b are constants.

The reason for this is obvious. Any method of estimation must be such that when we estimate θ we must at the same time be estimating θ^2, θ^3 ; in fact all explicit functions of θ . Now, in general, to put θ equal to the mean or expected value of T is not the same as putting $f(\theta)$ equal to $f\{E(T)\}$; in other words

$$f\{E(T)\} \neq E\{f(T)\}$$

(cf. our criticisms of Steffensen's self-induced "paradoxes.") But from out of all the mutually inconsistent estimates that we can obtain by putting various explicit functions of θ equal to the expectations of the corresponding functions of the estimating function, it is possible to choose the one which has the minimum variance. If this is obtained by estimating for $F(\theta)$, we must obtain θ from this estimate. Later we will obtain a theorem which tells us in each case for what particular function of θ we should estimate.

The existence of this "basic" function of θ , as we will term it, also emerges in the method of maximum likelihood, in the fact that we can obtain the maximum of the likelihood function by differentiating $\log L$ with respect to θ or θ^2 or θ^3 , or

any function of θ to obtain the maximum. For instance to estimate for σ in the normal curve, we get

$$\frac{\partial}{\partial \sigma} \log \Phi = -\frac{n}{\sigma} + \frac{\Sigma (x_r - m)^2}{\sigma^3},$$

$$\therefore \hat{\sigma}^2 = \frac{1}{n} \Sigma (x_r - m)^2 .$$

If we estimate for σ^2, we get

$$\frac{\partial}{\partial \sigma^2} \log \Phi = -\frac{n}{2\sigma^2} + \frac{1}{2} \frac{\Sigma (x_r - m)^2}{\sigma^4},$$

$$\therefore \hat{\sigma}^2 = \frac{1}{n} \Sigma (x_r - m)^2,$$

exactly as before; and so on.

3·09 The Non-existence of Solutions

The second important point is brought out by example V. Here the nature of the difficulty differs from that in example IV. In example IV we were simply estimating for the wrong function of θ, but in the case of the Cauchy distribution it appears that there is no function of θ for which we can estimate and obtain a solution. (It is well known in the Calculus of Variations that many plausibly posed minimal problems have in fact no analytic solution.) Hence there must be something fundamentally different about a curve such as the Cauchy curve by comparison with, say, the normal curve. This point will be elaborated in the discussion of sufficient statistics. What is of importance here is that Fisher, in his manner of presenting the problem of estimation has passed rather quickly over an unelucidated difficulty by saying that we find $\hat{\theta}$ in <u>every</u> case by putting

$$\frac{\partial}{\partial\theta}\log\Phi = 0 \ ;$$

that is to say, for example, that we obtain $\hat{m}$ in the Cauchy distribution by solving

$$\sum_{r=1}^{n} \frac{x_r - \hat{m}}{1+(x_r-\hat{m})^2} = 0.$$

One final point of interest. We have developed our method, and, in certain cases at any rate, justified Fisher's method and results, without any recourse whatever to the notion of likelihood. Nor in our fundamental approach do we regard θ as being a disposable parameter, as does Fisher. We state the problem as follows:

"Choose a function T of the observations, such that the expectation of T taken over all possible samples of n should be equal to θ .

Then out of all such functions take that which has the minimum variance."

The fact that $\frac{\partial}{\partial\theta}\log\Phi$ appears in our final result is due, not to our fundamental approach, but to the devices employed in the Calculus of Variations.

If $\frac{\partial}{\partial\theta}\log\Phi$ takes the form

$$\frac{2}{\lambda(\theta)}\{\theta - f(x_1,\cdots,x_n),$$

then our solution is

$$f(x_1,\cdots,x_n) = \theta - \frac{\lambda(\theta)}{2}\frac{\partial}{\partial\theta}\log\Phi$$

or is in effect given by

$$\hat{\theta} = f(x_1,\cdots,x_n),$$

just as with the method of maximum likelihood.

The duality or parallelism apparent in these two fundamentally different approaches is worthy of fuller investigation. For instance we find the same duality in the two different approaches to the method of least squares; and doubtless some similar kind of parallelism, perhaps limited, exists between fiducial probability and the confidence intervals of J.Neyman.

3·10 When is a Solution Possible?

It will be of interest to find some of the possible types of curve which will yield a solution.

We wish to have

$$\frac{\partial}{\partial\theta}\log\Phi = \frac{2}{\lambda(\theta)}\{\theta - T(x_1, \cdots, x_n)\}$$

i.e. $\frac{\partial}{\partial\theta}\log\Phi$ must break up into the form

$$\mu(\theta)\{\theta - T\}$$, say.

It is very difficult to say what kind of function Φ will yield such a result; but if we confine our attention to analytic differentiable functions, we can enumerate some of the possible cases by putting various functions of θ for $\mu(\theta)$, and performing the integration. The integration is possible in simple terms provided $\mu(\theta)$ possesses one of the following forms: a polynomial in θ ; a polynomial in $\frac{1}{\theta}$; a polynomial in both θ and $\frac{1}{\theta}$; an exponential function of a linear function of θ , including the case of imaginary coefficients, i.e. sine and cosine of a linear function; a product of a polynomial in θ and an exponential of these types; the logarithm of a linear function

of θ ; arc sine or arc tangent function. Other types may be yielded by other functions $\mu(\theta)$, but these are not finitely expressible in rational functions and the standard transcendentals.

3·11· Examples

I $$\mu(\theta) = n e^{-\theta} ;$$

$$\frac{\partial}{\partial\theta} \log \Phi = n\theta e^{-\theta} - n T e^{-\theta},$$

$$\log \Phi = -n\theta e^{-\theta} + n e^{-\theta} + n T e^{-\theta} + \text{const.},$$

$$\therefore \quad \Phi = c^n e^{n(T-\theta-1)e^{-\theta}},$$

a curious double exponential.

Further complications are of course introduced by the necessity of specifying T , which is to be some symmetrical function of the x_r . If, say, $T = \frac{\Sigma x}{n}$, we get for the probability function

$$\phi = c\, e^{(x-\theta-1)e^{-\theta}}.$$

The double exponential type of curve is not an uncommon one, as double exponentials appear in the distribution of the range as well as in the distributions of the largest and smallest values in a sample from a normal population.

II $$\mu(\theta) = \frac{n}{1+\theta^2}$$

$$\frac{\partial}{\partial\theta} \log \Phi = n(\theta - T)/(1+\theta^2),$$

$$\therefore \log \Phi = \frac{n}{2} \log(1+\theta^2) - n T \arctan\theta + \text{const.}$$

$$\therefore \quad \Phi = c^n (1+\theta^2)^{\frac{n}{2}} e^{-n\bar{T} \arctan \theta} \;;$$

$$\therefore \quad \varphi = c\sqrt{1+\theta^2}\; e^{-x \arctan \theta} \;,$$

if, say, $\bar{T} = \frac{1}{n}\Sigma x$.

III $\qquad \mu(\theta) = -n$

$$\log \Phi = -\int n(\theta - \bar{T})\, d\theta + \text{const.}$$

$$= -\tfrac{1}{2} n\theta^2 + n\bar{T}\theta + \text{const.}$$

$$\therefore \quad \Phi = c^n e^{-\frac{1}{2}n(\theta^2 - 2\bar{T}\theta)}$$

If $\bar{T} = \frac{\Sigma x}{n}$, say, this leads to

$$\varphi = \kappa e^{-\frac{1}{2}(x-\theta)^2} \;,$$

since the constant can involve x but not θ.

IV $\qquad \mu(\theta) = -n\theta$

$$\frac{\partial}{\partial \theta} \log \Phi = -n(\theta^2 - \bar{T}\theta) \;,$$

$$\therefore \quad \log \Phi = -n\left(\frac{\theta^3}{3} - \frac{\bar{T}\theta^2}{2}\right) + \text{const.}$$

This does not lead to a known form of probability function. Neither do the polynomials of higher degree in θ.

V $\qquad \mu(\theta) = -\frac{n}{\theta}$

$$\frac{\partial}{\partial \theta} \log \Phi = -n + \frac{n\bar{T}}{\theta}$$

$$\therefore \quad \log \Phi = -n\theta + n\bar{T} \log \theta + \text{const.}$$

$$\therefore \Phi = K^n e^{-n\theta} \theta^{nT}$$

If $T = \frac{1}{n}\Sigma x$, say, this gives

$$\phi = K e^{-\theta} \theta^{x},$$

or, possibly, since the constant can involve x ,

$$\phi = e^{-\theta} \frac{\theta^x}{x!}$$

the Poisson function with θ as mean.

VI $$\mu(\theta) = -\frac{n}{\theta^2}$$

$$\frac{\partial}{\partial\theta} \log \Phi = -\frac{n}{\theta} + \frac{nT}{\theta^2}$$

$$\therefore \log \Phi = -n \log \theta - \frac{nT}{\theta} + \text{const.}$$

$$\therefore \Phi = K^n \theta^{-n} e^{-\frac{nT}{\theta}}$$

If $T = \frac{1}{n}\Sigma x$,

$$\phi = K \theta^{-1} e^{-\frac{2x}{\theta}},$$

one of the Pearsonian curves.

VII $$\mu(\theta) = -\frac{n}{2\theta^2}$$

$$\frac{\partial}{\partial\theta} \log \Phi = -\frac{n}{2\theta} + \frac{nT}{2\theta^2},$$

$$\therefore \log \Phi = -\frac{n}{2} \log \theta - \frac{nT}{2\theta} + \text{const},$$

$$\therefore \Phi = K \theta^{-\frac{n}{2}} e^{-\frac{nT}{2\theta}}.$$

If $T = \frac{1}{n}\Sigma x^2$,

$$\phi = c\,\theta^{-\frac{1}{2}} e^{-\frac{n\bar{x}^2}{2\theta}}$$

$$= c\,\sigma^{-1} e^{-\frac{n\bar{x}^2}{2\sigma^2}}, \quad \text{if} \quad \theta = \sigma^2.$$

$\mu(\theta) = -\frac{n}{\theta^3}$ leads to a function of the type

$$\phi = \kappa e^{-\frac{1}{2}\left(\frac{n}{\theta} - \frac{1}{n}\right)^2}.$$

Polynomials other than those of the second degree or lower in $\frac{1}{\theta}$ do not lead to known forms of probability functions. The types obtained are possible, but unusual and sometimes bizarre.

$\mu(\theta) = \log \theta$ leads to an expression of the form

$$\Phi = \kappa \theta^{\left(\frac{1}{2}\theta^2 - T\theta\right)} e^{-\left(\frac{\theta^2}{4} - T\theta\right)}.$$

. . . . .

$$\mu(\theta) = \frac{n}{1-\theta^2}$$

$$\frac{\partial}{\partial\theta} \log \Phi = n(\theta - T)/(1-\theta^2).$$

$$\log \Phi = -\frac{n}{2}\log(1-\theta^2) - \frac{nT}{2}\log\frac{1+\theta}{1-\theta} + \text{const.},$$

$$\therefore \Phi = K(1-\theta^2)^{-\frac{n}{2}} \times \left(\frac{1+\theta}{1-\theta}\right)^{-\frac{nT}{2}}$$

If $T = \frac{1}{n}\Sigma x$,

$$\phi = c(1-\theta^2)^{-\frac{1}{2}}\left(\frac{1+\theta}{1-\theta}\right)^{-\frac{n\bar{x}}{2}}$$

$$= c(1+\theta)^{-\frac{1+n\bar{x}}{2}}(1-\theta)^{-\frac{1-n\bar{x}}{2}}.$$

It would of course be possible to extend without limit the number of curves yielding solutions. The enumeration of such

curves by the above method is not a very satisfactory process, because of the arbitrary manner of selecting the function $T(x_1, \cdots, x_n)$ and also the constants introduced by integrating. Of course the constants may be evaluated in the final form of the distribution φ from the fact that the curve must enclose unit area with the x-axis. However, as we have seen, the constants themselves might be functions of x, and the choosing of these functions is often quite an arbitrary process. Then, again, in the expression

$$M(\theta)\{\theta - T\}$$

θ itself can, as we have seen, and, in fact, usually is a function of the corresponding parameter specifying a curve, as for instance in example VII where we put θ equal to σ^2 in the final form of φ in order to obtain the normal curve.

3·12. The Variance of our Estimate

We have

$$\int\int\cdots\int \Phi \cdot dx_1 \cdots dx_n = 1 \qquad \cdots (1)$$

$$\int\int\cdots\int (T-\theta)\cdot \Phi \cdot dx_1 \cdots dx_n = 0 \qquad \cdots (2)$$

$$V = \int\int\cdots\int (T-\theta)^2 \Phi \, dx_1 \cdots dx_n = \min \qquad \cdots (3)$$

V is a minimum when $T - \theta = -\frac{\lambda}{2}\Phi_\theta/\Phi$,

λ being a function of θ only.

Hence

$$V = \frac{\lambda^2}{4}\int\int\cdots\int \left(\frac{\Phi_\theta}{\Phi}\right)^2 \Phi \, dx_1 \cdots dx_n, \qquad \cdots (4)$$

where

$$\iiint \cdots \Phi_\theta \cdot dx_1 \cdots dx_n = 0 \qquad \cdots (5)$$

by (2).

Now

$$\frac{\partial^2}{\partial\theta^2} \log \Phi = \frac{1}{\Phi}\frac{\partial^2 \Phi}{\partial \theta^2} - \left(\frac{\Phi_\theta}{\Phi}\right)^2$$

$$\therefore \left(\frac{\Phi_\theta}{\Phi}\right)^2 \Phi = \frac{\partial^2 \Phi}{\partial \theta^2} - \frac{\partial^2}{\partial\theta^2} \log \Phi ,$$

and

$$\iiint \cdots \frac{\partial^2 \Phi}{\partial \theta^2} dx_1 \cdots dx_n = 0 ,$$

$$\therefore -V = \frac{\lambda^2}{4} \iiint \cdots \left(\frac{\partial^2}{\partial\theta^2} \log \Phi\right) \Phi \cdot dx_1 \cdots dx_n . \qquad \cdots (6)$$

3·13. Examples

Variance of mean of normal curve,

$$\phi = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-m)^2/\sigma^2} ,$$

$$\Phi = (2\pi)^{-\frac{1}{2}n} \sigma^{-n} e^{-\frac{1}{2}\Sigma(x_r - m)^2/\sigma^2} ,$$

$$\frac{\partial}{\partial\theta} \log \Phi = -\frac{1}{\sigma^2}(nm - \Sigma x)$$

$$\frac{\partial^2}{\partial\theta^2} \log \Phi = -\frac{n}{\sigma^2} \quad ; \quad \frac{\lambda}{2} = -\frac{\sigma^2}{n} .$$

$$-V = \iiint \cdots \frac{\sigma^4}{n^2}\left(-\frac{n}{\sigma^2}\right) \Phi \, dx_1 \cdots dx_n$$

$$= -\frac{\sigma^2}{n}$$

$$\therefore V = \frac{\sigma^2}{n} .$$

3·14 Variance of σ^2 in normal curve

$$\frac{\partial}{\partial\theta}\log\Phi = -\frac{n-1}{2\sigma^2} + \frac{1}{2}\frac{\Sigma(x_r-\bar{x})^2}{\sigma^4}, \qquad \theta=\sigma^2,$$

$$\frac{\partial^2}{\partial\theta^2}\log\Phi = \frac{n-1}{2\sigma^4} - \frac{\Sigma(x_r-\bar{x})^2}{\sigma^6}$$

and

$$\frac{\lambda}{2} = -\frac{2\sigma^4}{n-1}, \quad \therefore \frac{\lambda^2}{4} = \frac{4\sigma^8}{(n-1)^2};$$

$$\therefore -V = \iiint\cdots\left[\frac{2\sigma^4}{n-1} - \frac{4\sigma^2\cdot\Sigma(x_r-\bar{x})^2}{(n-1)^2}\right]\Phi\, dx_1\cdots dx_n$$

$$= \frac{2\sigma^4}{n-1} - \frac{4\sigma^2}{(n-1)^2}\iiint\cdots\Sigma(x_r-\bar{x})^2\cdot\Phi\cdot dx_1\cdots dx_n$$

But since

$$\iiint\cdots\Phi_{\sigma^2}\, dx_1\cdots dx_n = 0,$$

$$\iiint\cdots\frac{\Sigma(x_r-\bar{x})^2}{2\sigma^4}\cdot\Phi\cdot dx_1\cdots dx_n$$

$$= \iiint\cdots\frac{n-1}{2\sigma^2}\cdot\Phi\cdot dx_1\cdots dx_n .$$

$$\therefore -V = \frac{2\sigma^4}{n-1} - \frac{4\sigma^2}{(n-1)^2}(n-1)\sigma^2$$

$$= -\frac{2\sigma^4}{n-1}$$

$$\therefore V = \frac{2\sigma^4}{n-1}$$

3·15

Returning to equation 3·12·(6)',

$$-V = \frac{\lambda^2}{4} \iiint \cdots \left(\frac{\partial^2}{\partial\theta^2} \log \Phi\right) \Phi \, dx_1 \cdots dx_n$$

If $\frac{\partial}{\partial\theta} \log \Phi$ can be written in the form

$$-\frac{2}{\lambda}(T-\theta)$$

we must have that

$$\frac{\partial^2}{\partial\theta^2} \log \Phi = \frac{2}{\lambda} - (T-\theta)\frac{\partial}{\partial\theta}\left(\frac{2}{\lambda}\right);$$

$$\therefore \iiint \cdots \left(\frac{\partial^2}{\partial\theta^2} \log \Phi\right) \Phi \, dx_1 \cdots dx_n$$

$$= \iiint \cdots \frac{2}{\lambda} \cdot \Phi \cdot dx_1 \cdots dx_n = \frac{2}{\lambda}$$

since

$$\iiint \cdots (T-\theta) \cdot \Phi \cdot dx_1 \cdots dx_n = 0,$$

and $\frac{2}{\lambda}$ is independent of x.

$$\therefore -V = \frac{\lambda^2}{4} \iiint \cdots \frac{2}{\lambda} \cdot \Phi \cdot dx_1 \cdots dx_n$$

$$= \frac{\lambda}{2};$$

or

$$-\frac{1}{V} = \frac{2}{\lambda}$$

$$= \iiint \cdots \left(\frac{\partial^2}{\partial\theta^2} \log \Phi\right) \Phi \, dx_1 \cdots dx_n.$$

$$\therefore -\frac{1}{\sigma_T^2} = E'\left(\frac{\partial^2 L}{\partial\theta^2}\right), \quad L = \log \Phi,$$

in Fisher's notation.

Here we have Fisher's well known formula holding accurately, without any approximations, and without any assumptions as to the magnitude of n or the sampling distribution of T. It must be noted that this formula holds accurately only provided $\frac{\partial}{\partial\theta}\log\Phi$ splits up into the form

$$\frac{2}{\lambda(\theta)}\{\theta - T\}.$$

3·16· Sufficient Statistics

When $\frac{\partial}{\partial\theta}\log\Phi$ splits up into the form

$$\frac{2}{\lambda(\theta)}\{\theta - T\},$$

(where, as has been noted, we can replace θ by a function of θ), T will be called a "sufficient statistic". We will now proceed to show that this conception of sufficiency is in conformity with that of Fisher.

B.O. Koopman, in a paper entitled "On Distributions Admitting a Sufficient Statistic," [Trans. Amer. Math. Soc., 39, (1936), p. 399,] solves the problem of finding what distributions of the form $\varphi(x\,|\,\theta_1,\cdots,\theta_k)$ will admit of the determining of sufficient statistics for the estimation of some or all of the parameters θ_i.

We will give an outline of the results for the estimation of a single parameter θ, although Koopman obtains them for parameters $\theta_1,---,\theta_n$.

We first give a mathematical definition of a sufficient statistic, the intuitive definition of which is a statistic which contains the whole of the relevant information concerning

which is supplied by the sample .

The distribution $\phi(x|\theta)$ leads to a sufficient statistic for the estimation of θ if

$$\frac{\Phi(x_1, \cdots, x_n|\theta)}{\Phi(x_1', \cdots, x_n'|\theta)} = \frac{\Phi(x_1, \cdots, x_n|\theta')}{\Phi(x_1', \cdots, x_n'|\theta')}$$

is implied by

$$T(x_1, \cdots, x_n) = T(x_1', \cdots, x_n'),$$

where $T(x_1, \cdots, x_n)$ is the statistic in question, and

$$\Phi(x_1, \cdots, x_n|\theta)\, dx_1 \cdots dx_n$$

is the compound probability of observations.

For example, if

$$\phi = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2},$$

$$\Phi = (2\pi)^{-\frac{1}{2}n} e^{-\frac{1}{2}\Sigma(x_r-\theta)^2}.$$

$$\frac{\Phi(x|\theta)}{\Phi(x|\theta')} = \frac{e^{-\frac{1}{2}\Sigma(x_r-\theta)^2}}{e^{-\frac{1}{2}\Sigma(x_r-\theta')^2}}$$

$$= e^{-\frac{1}{2}\Sigma\{(\theta^2-\theta'^2)-2x_r(\theta-\theta')\}};$$

$$\frac{\Phi(x'|\theta)}{\Phi(x'|\theta')} = e^{-\frac{1}{2}\Sigma\{(\theta^2-\theta'^2)-2x_r'(\theta-\theta')\}}.$$

Here $T(x_1, \cdots, x_n)$, a sufficient statistic for the estimation of θ is $\frac{1}{n}\Sigma x_r$.

Hence

$$T(x_1, - - -, x_n) = T(x_1', - - -, x_n')$$

means that

$$\frac{1}{n} \Sigma x_r = \frac{1}{n} \Sigma x_r'$$

and this implies that

$$\frac{\Phi(x|\theta)}{\Phi(x|\theta')} = \frac{\Phi(x'|\theta)}{\Phi(x'|\theta')} .$$

Koopman obtains the following theorems:

I If $\phi(x|\theta)$ be analytic and non-zero at each point in some sub-set of the region $I \times R - T$, where R is the axis of reals, $x_1, - - - - -, x_n$, and I the region, (one-dimensional,) containing θ;

and if $T(x_1, - -, x_n)$ be continuous in R^n, then a necessary condition that $T(x_1, - -, x_n)$ is a sufficient statistic for the estimation of θ is that, at each given point (a, b) of $I \times R - T$ a neighbourhood $N_{ab} = i_{ab} \times r_{ab} \subset R - T$ exist, where

$$i_{ab} : |\theta - a| < h$$

$$r_{ab} : |x - b| < h'$$

such that

$$\phi(x|\theta) = \exp b \left[\Theta_1 X_1 + \Theta + X \right] \ldots\ldots (1)$$

Θ_1 and Θ are real, single-valued, analytic functions of θ in i_{ab};

and X_1 and X are real, single-valued, analyticfunctions of x in r_{ab}.

Moreover, it follows that

$$\lambda_1 = V\{T(x_1, \cdots, x_n)\}, \quad \cdots\cdots (2)$$

where V is a single-valued function of $T(x)$.

II A sufficient condition is that (1) and (2) of theorem I should hold in the sub-set $R' \times I$ of (x, θ), where $R = R' + R''$, ($R'R'' = 0$), and $\varphi(x|\theta)$ is zero for all points in $R'' \times I$; and also, in $R' \times I$,

$$\int_R \varphi(x|\theta) = 1$$

III If the equation (1) of theorem I holds subject to $\varphi(x|\theta) = 0$ in $R'' \times I$; and if $\Phi(x_1, \cdots, x_n|\theta)$ has a unique maximum, $\hat{\theta}$, in I, ($\frac{\partial \Phi}{\partial \theta}$ existing and non-zero at $\hat{\theta}$,) for each $(x_1, \cdots, x_n)$ in $(R')^n$; then $\hat{\theta}$ is a sufficient statistic for the estimation of θ.

This theorem is equivalent to Fisher's theorem that, if a sufficient statistic exists, the method of maximum likelihood yields that statistic.

The importance of these theorems is that they enable us to say whether, for any distribution $\Phi(x|\theta_1, \cdots, \theta_n)$ it is possible to find a sufficient statistic $T_i(x_1, \cdots, x_n)$ for the estimation of θ_i, assuming that the θ_j, ($j \neq i$), are all known.

Koopman gives a general statement and proof of the theorems for the estimation of one or more of the θ_i. In this case the one-dimensional region I becomes a k-dimensional region Ω. Equation (1) of theorem I becomes

$$\phi(x \mid \theta_1, \cdots, \theta_n) = \exp\left[\sum_{i=1}^{l} \Theta_i X_i + \Theta + X\right],$$

where $l = 0$ implies that all the functions $\Theta_i X_i$ are missing.

3·11· If, now, we return to the equation

$$\frac{\partial}{\partial \theta} \log \Phi = \frac{2}{\lambda(\theta)} \{\theta - T\},$$

we have

$$\log \Phi = \int \frac{2}{\lambda(\theta)} (\theta - T) d\theta + \text{const}$$

where the constant can be a function of x, independent of θ.

Hence

$$\log \Phi = \mu(\theta)(\theta - T) - \int \mu(\theta) d\theta + \text{const.}$$

where

$$\mu(\theta) = \int_0^{\theta} \frac{2}{\lambda(\theta)} d\theta .$$

Hence

$$\log \Phi = \mu(\theta)(\theta - T) + \mu_1(\theta) + \text{const.}, \text{ say.}$$

i,e, is of the form

$$\left[F_1(\theta) f_{\Sigma_1}(x) + n F_2(\theta) + f_{\Sigma_2}(x)\right];$$

and as Φ is the product of n elemental functions in $x_1, \cdots, x_n$ respectively, it follows that the distribution ϕ must be of the form

$$\phi(x \mid \theta) = \exp\left[F_1(\theta) f_1(x) + F_2(\theta) + f_2(x)\right]$$

Hence if $\frac{\partial}{\partial \theta} \log \Phi$ takes the form $\frac{2}{\lambda(\theta)} \{\theta - T\}$, T will be a sufficient statistic. This means that every statistic that conforms to the condition of unbiassed minimum variance is a

sufficient statistic.

3·18. We have seen previously that there is in general only one function of θ for which we can estimate and obtain a solution by the method of consistent minimum variance. If that function is $t(\theta)$, we must estimate for $t(\theta)$ and solve if possible for θ. This is of extreme importance, because, as we shall see, there have been methods put forward which yield inconsistent results according as to whether we apply these methods to the estimation of θ, θ^2, or, generally, $F(\theta)$.

We can now find the particular function of θ for which we must estimate.

For, let that function be $t(\theta)$:

Then if

$$\phi(x|\theta) = \exp\left[F_1 f_1 + F_2 + f_2\right]$$

capital letters on the right hand side denoting functions of θ alone, and small letters functions of x alone, we have

$$\Phi = \exp\left[F_1(\Sigma f_1) + n F_2 + \Sigma f_2\right]$$

$$\therefore \frac{\partial}{\partial t}\log\Phi = \frac{dF_1}{dt}(\Sigma f_1) + n\frac{dF_2}{dt}$$

$$= n\frac{dF_1}{dt}\left\{\frac{\Sigma f_1}{n} + \frac{dF_2}{dF_1}\right\}$$

Hence if

$$t = -\frac{dF_2}{dF_1} = -\frac{dF_2}{d\theta}\Big/\frac{dF_1}{d\theta}$$

we get

$$\frac{\partial}{\partial t}\log\Phi = \mu(t)\left\{\frac{\Sigma f_1}{n} - t\right\},$$

which is of the required form.

Hence we estimate for $-\frac{dF_2}{dF_1}$.

3·19 Examples

I Mean of normal population.

$$\phi(x|m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-m)^2}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2 - 2mx + m^2)}$$

$$F_1 = m \quad ; \quad F_2 = -\tfrac{1}{2}m^2 .$$

$$\therefore -\frac{dF_2}{dF_1} = m .$$

Hence we estimate for m.

II Variance of normal population.

$$\phi(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2} - \log\sigma}$$

$$F_1 = -\frac{1}{2\sigma^2} \quad ; \quad F_2 = -\log\sigma$$

$$-\frac{dF_2}{dF_1} = -\frac{dF_2}{d\sigma} \Big/ \frac{dF_1}{d\sigma} = \frac{1}{\sigma} \Big/ \frac{1}{\sigma^3}$$

$$= \sigma^2$$

Hence we must estimate for σ^2.

3·20 The Variance of $\hat{t}$

We have also seen that if $\frac{\partial}{\partial\theta}\log\Phi$ can be written in the form $\frac{2}{\lambda(\theta)}\{\theta - T\}$, the variance V_T of T is given by

$$V_T = -\frac{\lambda}{2} .$$

Hence our estimate of $t(\theta)$ has variance given by

$$\frac{1}{V_{\hat{t}}} = n \frac{dF_1}{dt} = n \frac{dF_1}{d\theta} \Big/ \frac{dt}{d\theta}$$

and, since

$$t = -\frac{dF_2}{dF_1}$$

$$-\frac{1}{V_{\hat{t}}} = n \frac{dF_1}{d\theta} \Big/ \left(\frac{d^2F_2}{dF_1^2} \cdot \frac{dF_1}{d\theta} \right)$$

$$\therefore V_{\hat{t}} = -\frac{1}{n} \frac{d^2F_2}{dF_1^2}.$$

For example to find the variance of $\hat{\sigma}^2$ in the normal distribution, (remembering that n is replaced by $n-1$,) we have

$$\frac{dF_2}{dF_1} = -\sigma^2$$

$$\frac{d^2F_2}{dF_1^2} = -2\sigma \Big/ \frac{dF_1}{d\sigma}$$

$$= -2\sigma \Big/ \frac{1}{\sigma^3} = -2\sigma^4$$

$$\therefore V_{\hat{\sigma}^2} = \frac{2\sigma^4}{n-1}$$

3·21· We can now show that the only Pearsonian distribution for which the mean is a sufficient statistic for locating the curve is the normal distribution. For, since the variance of the mean of any distribution is $\frac{\sigma^2}{n}$, we must have, setting $\sigma^2 = 1$,

$$\frac{\partial}{\partial\theta}\log\Phi = -n\left(\theta - \frac{\Sigma x}{n}\right)$$

$$= -(n\theta - \Sigma x)$$

$$\therefore \frac{\partial}{\partial\theta}\log\varphi = -(\theta - x)$$

$$\therefore \log\varphi = -\tfrac{1}{2}(x-\theta)^2 + \text{const.}$$

$$\therefore \varphi = c\,e^{-\frac{1}{2}(x-\theta)^2}.$$

Since

$$\int_{-\infty}^{\infty} e^{-z^2}\,dz = \sqrt{\pi}$$

$$\varphi = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(x-\theta)^2}.$$

The only arbitrary step in this process has been the choice of the constant, for it must be remembered that the constant can involve x but not θ. Hence the normal curve is not the only one we could arrive at as our final form, but a glance at the Pearsonian curves will show that we cannot arrive at any Pearsonian forms other than the normal, no matter how we choose our constant.

3·22. R.A. Fisher and Bayes' Theorem (1921 paper)

The chance of obtaining x successes and y failures in a sample of n from an infinite population containing a proportion p of successes is

$$\frac{n!}{x!\,y!}\,p^x(1-p)^y \qquad \text{------ (1)}$$

The optimum value, $\hat{p}$, of p is that which, according to the theory of maximum likelihood, makes (1) a maximum.

For this, differentiating the logarithm of (1) with respect to p, we get

$$\frac{x}{\hat{p}} = \frac{y}{1-\hat{p}} = \frac{n-x}{1-\hat{p}}$$

$$\therefore \hat{p} = \frac{x}{n}$$

Bayes' Theorem discusses the accuracy of this determination by enquiring: "When we know $\hat{p}$, what is the probability that p lies in the range dp?" Bayes assumes this probability to be simply dp.

"After the selection effected by observing the sample," says Fisher, "the probability is clearly proportional to $p^x(1-p)^y dp$. After giving this solution, ... Bayes adds a scholium the purport of which would seem to be that in the absence of all knowledge save that supplied by the sample, it is reasonable to assume that particular a priori distribution of p."

Fisher points out the arbitrary nature of this assumption and claims that the solution is not even unique; and proves his assertion thus:

Instead of considering p, let us measure probability on a different scale. For instance if

$$\sin\theta = 2p-1,$$

then θ measures probability just as well as p itself. The chance of x successes and y failures is now

$$\Phi = \frac{n!}{2^n x!y!}(1+\sin\theta)^x(1-\sin\theta)^y$$

$$\therefore \frac{\partial}{\partial\theta}\log\Phi = \frac{x\cos\theta}{1+\sin\theta} - \frac{y\cos\theta}{1-\sin\theta} = 0;$$

$$\therefore \sin\theta = \frac{x-y}{n}$$

i.e. $2p-1 = \frac{2x-n}{n}$ $\quad\therefore p = \frac{x}{n}$, as before.

If, says Fisher, we are now to take as our a priori assumption that θ is equally likely to lie in all equal ranges $d\theta$, then our a priori probability will be $\frac{d\theta}{\pi}$, and that after making the observations will be proportional to

$$(1+\sin\theta)^x\,(1-\sin\theta)^y\,d\theta\,;$$

but if we interpret this in terms of p , we get

$$p^x(1-p)^y\,\frac{dp}{\sqrt{p(1-p)}}$$

$$= p^{x-\frac{1}{2}}(1-p)^{y-\frac{1}{2}}\,dp,$$

a result inconsistent with that obtained previously.

3·23· Actually, whilst Fisher says that θ serves to measure probability just as well as p , what he has been estimating for is not θ but $\sin\theta$, as the following will show:

$$\log\left[\frac{n!}{x!\,y!}\,p^x(1-p)^{n-x}\right]$$

$$= \text{const.} + x\log p - x\log(1-p) + n\log(1-p)$$

$$= \text{const} + x\{\log p - \log(1-p)\} + n\log(1-p).$$

Here

$$F_1 = \log p - \log(1-p) \quad ; \quad \frac{dF_1}{dp} = \frac{1}{p(1-p)} \; ;$$

$$F_2 = n\log(1-p) \quad ; \quad \frac{dF_2}{dp} = -\frac{n}{1-p} \; ;$$

$$-\frac{dF_2}{dF_1} = np.$$

Hence we estimate for n/p or, what amounts to the same thing, for p.

$$\log\left[\frac{n!}{2^n x!\,y!}(1+\sin\theta)^x(1-\sin\theta)^{n-x}\right]$$

$$= \text{const.} + x\{\log(1+\sin\theta) - \log(1-\sin\theta)\} + n\log(1-\sin\theta).$$

Here

$$F_1 = \log(1+\sin\theta) - \log(1-\sin\theta),$$

$$\frac{dF_1}{d\theta} = \frac{2}{\cos\theta},$$

$$\frac{dF_2}{d\theta} = \frac{d}{d\theta}\{n\log(1-\sin\theta)\}$$

$$= -\frac{n\cos\theta}{1-\sin\theta}.$$

$$\therefore -\frac{dF_2}{dF_1} = 2n(1+\sin\theta).$$

Hence we estimate for $\sin\theta$. But this is precisely the same thing as estimating for p itself. In other words we cannot estimate directly for θ, but we must estimate for $\sin\theta$, and take our estimate of θ from this.

3·24. Consequently it is rather difficult to see what the theory of estimation has to do with the proof or the disproof of Bayes' Theorem. If Fisher's aim was to show that different concepts of probability scale involve different "a priori" assumptions, (just as different geometries involve different axioms,) then why have reference to the theory of estimation to prove that? And, in fact, we have seen that Fisher has actually been operating in the

same scale right up to the point where he obtains his expression for $\sin\theta$. What Fisher has proved is not the "fallacy" of Bayes' Theorem, but the fact that we do, at least, need to be very careful as to how we use that theorem; that is if we feel the need to use it at all.

3·25· Comparison with Maximum Likelihood

Analogy with Least Squares

In the simplest case of least squares we desire to find the "best" value $\hat{x}$ of a variate of which we have n independent measurements x_i, assuming that $\hat{x}$ will be some linear combination of the x_i .

The original method depended upon postulates similar to those of maximum likelihood in the theory of estimation: namely maximum compound probability of observations and normal distribution of errors. That is to say we have to maximise

$$c \prod \left[\sigma_i^{-1} (2\pi)^{-\frac{1}{2}} \exp\left\{ -\tfrac{1}{2} (x_i - \hat{x}_i)^2 / \sigma_i^2 \right\} \right]$$

i.e.

$$\frac{\partial}{\partial \hat{x}_i} \sum \omega_i (x_i - \hat{x}_i)^2 = 0 ,$$

where ω_i = weight of observation x_i .

This is the same thing as minimising the sum of squared residuals weighted, and gives

$$\hat{x} = \frac{\sum \omega_i x_i}{\sum \omega_i} .$$

But it appears that this result has a much wider validity than was assumed in postulating a normal distribution of errors; for it is possible to arrive at precisely the same result by

postulating

(a) consistency of estimate; and

(b) minimum error variance;

without <u>any</u> assumptions as to the actual distribution of errors.

Let

$$\hat{x} = \Sigma b_i x_i = b'x$$

b' and x denoting row and column vectors in the ordinary sense. Now in order that our linear combination should be consistent, it must be such that, in the event of all the x_i being equal, we must get

$$\hat{x} = x_i \ ;$$

i.e.

$$\Sigma b_i = 1 .$$

Hence the linear combination is a weighted mean.

Now the variance of $b'x$ is $\Sigma b_i^2 \sigma_i^2$

$$= b'Vb ,$$

where V is the "variance Matrix" $[v_{ij}]$; v_{ij} being the product variances $\rho_{ij}\sigma_i\sigma_j$, and v_{ii} the variances σ_i^2 .

Now the postulate (b) of minimum error variance means that we have to find the minimum of $\Sigma b_i^2\sigma_i^2$, subject to the condition $\Sigma b_i = 1$. Introducing a Lagrange multiplier λ , we consider

$$S^2 = \tfrac{1}{2}\Sigma b_i^2\sigma_i^2 - \lambda(\Sigma b_i - 1) .$$

$\frac{\partial S^2}{\partial b_i} = 0$ gives

$$b_i\sigma_i^2 = \lambda \quad , \quad i = 1, 2, \cdots, n .$$

i.e.

$$b_i \propto 1/\sigma_i^2 \propto w_i ,$$

$$\therefore \ b_i = \frac{w_i}{\Sigma w_i} \ , \text{ since } \ \Sigma b_i = 1 ,$$

and

$$\hat{x} = \frac{\Sigma w_i x_i}{\Sigma w_i} .$$

From this example it is easily seen that the relationship that exists between these two methods is exactly the relationship that exists between the general method of estimation that we have outlined and Fisher's Method of Maximum Likelihood. We also see that, in the cases when our method coincides with Fisher's, namely when $\frac{\partial}{\partial\theta} \log \Phi$ can be written in the form

$$\mu(\theta)\{F(\theta) - T\}$$

$\mu(\theta)$ and $F(\theta)$ being functions of θ only, and T not involving θ, the maximum likelihood solution does not require the postulates of large n and normal sampling distribution of the estimate.

3·26 The Distribution of Sufficient Statistics

If we are estimating for $F_1(\theta)$, and if our estimate of F_1 is a sufficient statistic, then we must have

$$\frac{\partial}{\partial F_1} \log \Phi = F_2\{n F_1 - \Sigma f_1\},$$

where F_2 is a function of θ alone, and f_1 is a function of x alone;

$$\therefore \log \Phi = \int F_2 (n F_1 - \Sigma f_1) \, dF_1$$

$$= n \int F_1 F_2 \, dF_1 - \Sigma f_1 \int F_2 \, dF_1 + \Sigma f_2$$

where f_2 is a function of x alone.

Let

$$\int F_2 \, dF_1 = -\psi(\theta) ;$$

then

$$F_2 = -\frac{d\psi(\theta)}{dF_1}$$

and

$$n \int F_1 F_2 \, dF_1 = -n \int F_1 \, d\psi(\theta) .$$

Hence

$$\Phi = e^{-n \int F_1 \, d\psi(\theta) + \psi(\theta) \Sigma f_1 + \Sigma f_2}$$

i.e.

$$\Phi(\psi) = e^{-n \int F_1 \, d\psi + \psi \Sigma f_1 + \Sigma f_2}$$

$$= e^{-n F(\psi) + \psi \Sigma f_1 + \Sigma f_2} .$$

This is equivalent to the expression given by Fisher on page 294 of his paper "Two New Properties of Mathematical Likelihood," and the following proof is due to Fisher.

$$\Phi(\psi+it) = e^{-n F(\psi+it) + (\psi+it) \Sigma f_1 + \Sigma f_2}$$

$$\therefore \frac{\Phi(\psi)}{\Phi(\psi+it)} = e^{n F(\psi+it) - n F(\psi) - it \Sigma f_1} ,$$

$$= e^{-it S(X)} \; e^{n F(\psi+it) - F(\psi)} ,$$

where X is a function of the individual observations x namely f_1 .

The frequency function of X is

$$e^{-F(\psi)}\, e^{X\psi}\, e^{X_1}\, \frac{dx}{dX}\, dX,$$

where X_1 is a function of the observations, namely f_2.

Hence its characteristic function is

$$M(it) = e^{F(\psi+it) - F(\psi)},$$

whilst that of $S(X)$ is

$$\{M(it)\}^n.$$

The probability that $S(X)$ lies between S_0 and S_1 is

$$\int df = \frac{1}{2\pi}\int_0^\infty \frac{dt}{it}\left\{ e^{-iS_0t} M^n(it) - e^{iS_0t} M^n(-it) - e^{-iS_1t} M^n(it) + e^{iS_1t} M^n(-it)\right\}$$

$$= \frac{1}{2\pi}\int_0^\infty \frac{dt}{it}\left\{ \frac{\Phi(\psi, S_0)}{\Phi(\psi+it, S_0)} - \frac{\Phi(\psi, S_0)}{\Phi(\psi-it, S_0)} - \frac{\Phi(\psi, S_1)}{\Phi(\psi+it, S_1)} + \frac{\Phi(\psi, S_1)}{\Phi(\psi-it, S_1)}\right\},$$

S_1 and S_0 being the limits of the function $nF_1(T)$ (p.294)

3·27 Sufficient and Non-sufficient Statistics

We know that

$$\frac{\partial}{\partial\theta}\log\Phi(x|\theta) = \frac{\partial}{\partial\theta}\log\Phi(x|\hat\theta) + (\theta-\hat\theta)\frac{\partial^2}{\partial\theta^2}\log\Phi\{x|\hat\theta+K(\theta-\hat\theta)\},$$

$$0 < K < 1$$

For a sufficient statistic, the R.H.S. of the equation can be written accurately in the form

$$\frac{2}{\lambda(\theta)}(\theta-\hat{\theta})$$

i.e. the R.H.S. contains the factor $(\theta-\hat{\theta})$.

Hence, since the first term on the R.H.S. is independent of θ altogether, we must have

$$\frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) = 0$$

Accordingly, at least so far as sufficient statistics are concerned, we have verified Fisher's results without having recourse to the notion of likelihood.

Also

$$\frac{2}{\lambda(\theta)} = \frac{\partial^2}{\partial\theta^2}\log\Phi\left(x|\theta+K\cdot\overline{\theta-\hat{\theta}}\right);$$

and since, in practice, when $\frac{\partial}{\partial\theta}\log\Phi$ splits up into the form $\frac{2}{\lambda(\theta)}(\theta-\hat{\theta})$, we get $\hat{\theta}$ by putting $\theta=\hat{\theta}$ in $\frac{\partial}{\partial\theta}\log\Phi = 0$, we must get $\frac{2}{\lambda(\theta)}$, i.e. $-\frac{1}{V\hat{\theta}}$ from $\frac{\partial^2}{\partial\theta_2}\log\Phi(x|\theta)_{\theta=\hat{\theta}}$.

This is illustrated in the following examples.

3·28. Examples

I $$\varphi = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2},$$

$$\hat{\theta} = \frac{\Sigma x}{n} \quad ; \text{ i.e. } \Sigma x = n\hat{\theta}.$$

$$\frac{\partial}{\partial\theta}\log\Phi = \Sigma x - n\theta,$$

$$\frac{\partial^2}{\partial\theta^2}\log\Phi = -n\ ;$$

$$\therefore \frac{\partial^2}{\partial\theta^2} \log \Phi(x|\theta)_{\theta=\hat{\theta}} = -n$$

$$\therefore V_{\hat{\theta}} = \frac{1}{n}$$

or, in the usual notation,

$$V_{\hat{\theta}} = \frac{\sigma^2}{n}$$

II $\quad \phi = \sqrt{\frac{1}{2\pi\theta}}\, e^{-\frac{1}{2}\frac{x^2}{\theta}}$

$$\hat{\theta} = \frac{\Sigma x^2}{N}, \quad N = n-1 .$$

$$\therefore \Sigma x^2 = N\hat{\theta}$$

$$\frac{\partial^2}{\partial\theta^2} \log \Phi = \frac{N}{2}\theta^{-2} - \Sigma x^2 \cdot \theta^{-3}$$

$$\therefore \frac{\partial^2}{\partial\theta^2} \log \Phi(x|\theta)_{\theta=\hat{\theta}} = \frac{N}{2}\theta^{-2} - N\theta^{-2}$$

$$= -\frac{N}{2}\theta^{-2}$$

$$\therefore V_{\hat{\sigma}^2} = \frac{2\sigma^4}{n-1} .$$

III $\quad \Phi = \frac{n!}{x!\,r!}\,\theta^x (1-\theta)^{n-x}$,

$$\hat{\theta} = \frac{x}{n} .$$

$$\frac{\partial^2}{\partial\theta^2} \log \Phi = \frac{1}{(1-\theta)^2}\left(\frac{x}{\theta} - n\right) - \frac{1}{1-\theta}\cdot\frac{x}{\theta^2}$$

$$\therefore \frac{\partial^2}{\partial\theta^2} \log \Phi(x|\theta)_{\theta=\hat{\theta}} = -\frac{n}{\theta(1-\theta)} .$$

$$\therefore V_{\hat{\theta}} = \frac{\theta(1-\theta)}{n} .$$

3·29.
Since

$$\frac{\partial}{\partial\theta}\log\Phi(x|\theta) = \frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) + (\theta-\hat{\theta})\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\hat{\theta}) + \frac{1}{2!}(\theta-\hat{\theta})^3\frac{\partial^3}{\partial\theta^3}\log\Phi(x|\hat{\theta}) + \cdots$$

it follows that, apart from any considerations of sufficient statistics, the formula for

$$-\frac{1}{V_{\hat{\theta}}} = \frac{\partial^2}{\partial\theta^2}\log\Phi(x|\hat{\theta})$$

will hold accurately for distributions for which $\frac{\partial^3}{\partial\theta^3}\log\Phi(x|\theta)$ and higher derivatives are zero, and $\hat{\theta}$ is an unbiassed estimate making $\frac{\partial}{\partial\theta}\log\Phi$ zero.

For such distributions

$$\frac{\partial^2}{\partial\theta^2}\log\Phi = C_1,$$

where C_1 is a constant or zero or a function of x alone.

$$\therefore \frac{\partial}{\partial\theta}\log\Phi = C_1\theta + C_2$$

where C_2 is a constant or zero or a function of x alone.
If C_1 is a constant, not a function of x, then the distribution falls within the class of those yielding a sufficient statistic; i.e. the class for which

$$\frac{\partial}{\partial\theta}\log\Phi = \frac{2}{\lambda(\theta)}\{\theta - T\},$$

$\frac{2}{\lambda(\theta)}$ being independent of x.

If, on the other hand, C_1 is a function of x, and C_2 a constant (not a function of x) then we still have a distribution yielding a sufficient statistic, as $\frac{\partial}{\partial\theta}\log\Phi$ can be written in the form

$$c_2\theta\left\{\frac{1}{\theta}+\frac{c_1}{c_2}\right\}$$

indicating that a sufficient statistic can be found by estimating for $\frac{1}{\theta}$.

If c_1 and c_2 are both functions of x, then the distribution cannot yield a sufficient statistic.

3·30. We have seen that

$$\frac{\partial}{\partial\theta}\log\Phi = \frac{n}{\lambda(\theta)}\{\theta-\hat{\theta}\} = -\frac{1}{\sigma_{\hat{\theta}}^2}(\theta-\hat{\theta})$$

holds for a sufficient statistic without any assumptions as to the value of n or as to the ultimate distribution of $\hat{\theta}$.

$\hat{\theta}$ may be obtained by putting $\hat{\theta}$ equal to that value of θ which makes $\frac{\partial}{\partial\theta}\log\Phi$ equal to zero, just as in the method of maximum likelihood.

When $\frac{\partial}{\partial\theta}\log\Phi$ does not emerge in the form $\frac{n}{\lambda(\theta)}\{\theta-\hat{\theta}\}$, it means that in finite samples there is no unbiassed estimate which has minimum variance. In this case we must be content with the best "consistent" estimate, where "consistent" may be taken to mean "almost unbiassed."

3·31. Variance of Non-sufficient Statistics

$$\frac{\partial}{\partial\theta}\log\Phi = \frac{n}{\lambda(\theta)}\{\theta-\hat{\theta}\}$$

will now hold only in the limit when n is large, (infinite,) or as a first approximation when n is finite but sufficiently large.

Since

$$\frac{\partial}{\partial\theta}\log\Phi(x|\theta) = \frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) + (\theta-\hat{\theta})\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\theta'),$$

where θ' lies between θ and $\hat{\theta}$; we have, when n is large, since θ' tends to θ in the same way as $\hat{\theta}$,

$$\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\hat{\theta}) = \frac{2}{\lambda(\theta)}$$

$$= \iiint\cdots\left(\frac{\partial^2}{\partial\theta^2}\log\Phi\right)\Phi\, dx_1\cdots dx_n$$

provided the first term on the right is zero.

Hence

$$\sigma_{\hat{\theta}}^2 = -\frac{\lambda(\theta)}{2}$$

$$= 1\Big/\iiint\cdots\left(-\frac{\partial^2}{\partial\theta^2}\log\Phi\right)\Phi\, dx_1\cdots dx_n$$

$$= 1/ni$$

in Fisher's notation.

i is the amount of information in a single observation; i.e.

$$i = \int\left(-\frac{\partial^2}{\partial\theta^2}\log\varphi\right)\varphi\, dx .$$

The assumption also involves the normal distribution of $\hat{\theta}$ when n is large, since it may be shown independently that $\frac{\partial}{\partial\theta}\log\Phi$ tends to have a normal distribution.

The fact that the minimum variance condition holds shows that the maximum likelihood solution is a minimum variance solution for n finite, on the assumption that as n increases, the distribution of $\hat{\theta}$ approaches normality.

In large samples the maximum likelihood solution uses all the information available in the sample. In finite samples it

uses an amount of information which, in view of the minimum variance property, cannot be exceeded by that used by any other statistic.

It must be noted that the above results, at least so far as non-sufficient statistics are concerned, hold accurately only in the limit, or, at any rate, when n is sufficiently large to make

$$\frac{\partial}{\partial\theta}\log\Phi(x|\hat{\theta}) + (\theta-\hat{\theta})\frac{\partial^2}{\partial\theta^2}\log\Phi(x|\hat{\theta})$$

a good approximation to $\frac{\partial}{\partial\theta}\log\Phi(x|\theta)$.

Fisher has shown that the use of this approximation rejects a certain amount of the information when n is finite — but that the amount is measureable. He has also shown that part of it can be recovered by the use of "ancillary statistics," involving the use of the second and higher derivatives of $\log\Phi$; and that, theoretically at least, it can all be recovered by using the whole course of the likelihood function.

3·32 Simultaneous Estimation

For the simultaneous estimation of $\theta_1, \theta_2, \cdots, \theta_n$ in

$$\phi(x|\theta_1;\cdots;\theta_n)$$

we must have the following equations satisfied simultaneously:

$$T_1(x_1;\cdots;x_n) = \theta_1 + \frac{\lambda_1(\theta_1,\cdots,\theta_n)}{2}\frac{\partial}{\partial\theta_1}\log\Phi$$

$$\cdots\cdots\cdots\cdots \qquad \cdots\cdots\cdots\cdots$$

$$\cdots\cdots\cdots\cdots \qquad \cdots\cdots\cdots\cdots$$

$$T_n(x_1;\cdots;x_n) = \theta_n + \frac{\lambda_n(\theta_1;\cdots;\theta_n)}{2}\frac{\partial}{\partial\theta_n}\log\Phi$$

where $T_1, \cdots, T_n$ are functions of $x_1, \cdots, x_n$, only,
and $\lambda_1, \cdots, \lambda_n$ are functions of $\theta_1, \cdots, \theta_n$, only.

3·33· For example, let us try to estimate m and σ^2 simultaneously in

$$\phi(x|m,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-m)^2/\sigma^2},$$

$$\Phi = (2\pi)^{-\frac{1}{2}n} \sigma^{-n} e^{-\frac{1}{2}\Sigma(x_r-m)^2/\sigma^2}.$$

The conditions are:

(1)

$$\frac{nm}{\sigma^2} - \frac{\Sigma x_r}{\sigma^2} = \frac{2}{\lambda_1(m,\sigma^2)}\left[m - T_1\right]$$

(2)

$$-\frac{n}{2\sigma^2} + \frac{1}{2}\frac{\Sigma(x_r-m)^2}{\sigma^4} = \frac{2}{\lambda_2(m,\sigma)}\left[\sigma^2 - T_2\right].$$

From (1)

$$\lambda_1 = -\frac{2\sigma^2}{n}, \quad T_1 = \frac{1}{n}\Sigma x_r$$

(2), on the other hand cannot be solved, because m is inextricably involved with x on the left hand side. Hence the procedure is to accept the estimate of m and proceed as we did before, obtaining

$$T_2 = \frac{1}{n-1}\Sigma(x_r-m)^2, \quad m = \bar{x}.$$

The usual precept to divide $\Sigma(x_r-m)^2$ by n is useless, since m is unknown; also, as we have seen, it is incorrect when m is estimated.

Maximum likelihood takes no account of the difficulties

involved here; as, according to this method, we have

(1)

$$\frac{n\hat{m}}{\hat{\sigma}^2} - \frac{\Sigma x}{\hat{\sigma}^2} = 0 \quad , \quad \therefore \ \hat{m} = \frac{1}{n}\Sigma x \ ;$$

(2)

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2}\frac{\Sigma(x_r - \hat{m})^2}{\hat{\sigma}^4} = 0 \ , \quad \therefore \ \hat{\sigma}^2 = \frac{\Sigma(x-\bar{x})^2}{n} \ ,$$

the latter being incorrect, if the equations are solved as strict simultaneous equations.

3·34. To estimate r in

$$\phi(x, y \mid r) = \frac{1}{2\pi\sqrt{1-r^2}}\, e^{-\frac{1}{2}(x^2 - 2rxy + y^2)\frac{1}{1-r^2}} .$$

Here we are assuming that we have located the curve. Hence, just as in the case of estimating σ^2 in $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}$, we operate in terms of n but in our final result replace n by $n-1$.

We have

$$\Phi = (2\pi)^{-n}(1-r^2)^{-\frac{1}{2}n}\, e^{-\frac{1}{2}\Sigma(x_i^2 - 2rx_iy_i + y_i^2)\frac{1}{1-r^2}} ,$$

$$\frac{1}{n}\Sigma x^2 = 1 \quad , \quad \frac{1}{n}\Sigma y^2 = 1 .$$

Hence

$$\Phi = (2\pi)^{-n}(1-r^2)^{-\frac{1}{2}n}\, e^{\Sigma(rxy - n)\frac{1}{1-r^2}}$$

$$\therefore \frac{\partial}{\partial\theta}\log\Phi = -rn(1+r^2)/(1-r^2)^2 + \Sigma xy\,(1+r^2)/(1-r^2)^2$$

$$= \frac{2}{\lambda(r)}\left[r - T^2\right] .$$

$$\frac{\lambda}{2} = \frac{(1-r^2)^2}{n(1+r^2)}$$

gives

$$T = \frac{\Sigma xy}{n}$$

or

$$\hat{r} = \frac{\Sigma xy}{n-1}$$

where $\hat{r}$ is the corrected estimate.

3-35. To estimate simultaneously σ_x^2, σ_y^2 and r in

$$\phi(x,y|\sigma_x^2,\sigma_y^2,r) = \frac{1}{2\pi\sqrt{1-r^2}}\frac{1}{\sigma_x\sigma_y}\, e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} - 2r\frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)\frac{1}{1-r^2}}$$

$$\Phi = (2\pi)^{-\frac{1}{2}n}(\sigma_x\sigma_y)^{-n}(1-r^2)^{-\frac{1}{2}n}\, e^{-\frac{1}{2}\left(\frac{\Sigma x^2}{\sigma_x^2} - 2r\frac{\Sigma xy}{\sigma_x\sigma_y} + \frac{\Sigma y^2}{\sigma_y^2}\right)\frac{1}{1-r^2}}.$$

$$\frac{\partial}{\partial\sigma_x^2}\log\Phi = -\frac{n}{2}\sigma_x^{-2} + \frac{1}{2}\left(\Sigma x^2\cdot\sigma_x^{-4} - r\,\Sigma xy\cdot\sigma_x^{-3}\sigma_y^{-1}\right)\frac{1}{1-r^2}$$

$$\therefore T_1 = \sigma_x^2 - \frac{\lambda_1}{2}\left[\frac{n}{2}\sigma_x^{-2} - \frac{1}{2}\left(\Sigma x^2\cdot\sigma_x^{-4} - r\,\Sigma xy\cdot\sigma_x^{-3}\sigma_y^{-1}\right)\frac{1}{1-r^2}\right].$$

Similarly

$$T_2 = \sigma_y^2 - \frac{\lambda_2}{2}\left[\frac{n}{2}\sigma_y^{-2} - \frac{1}{2}\left(\Sigma y^2\cdot\sigma_y^{-4} - r\,\Sigma xy\cdot\sigma_y^{-3}\sigma_x^{-1}\right)\frac{1}{1-r^2}\right],$$

$$T_3 = r + \frac{\lambda_3}{2}\left[\frac{rn}{1-r^2} - \left(\frac{\Sigma x^2}{\sigma_x^2} - \frac{1+r^2}{r}\cdot\frac{\Sigma xy}{\sigma_x\sigma_y} + \frac{\Sigma y^2}{\sigma_y^2}\right)\frac{r}{(1-r^2)^2}\right].$$

Writing

$$\frac{\Sigma x^2}{\sigma_x^2} = X, \quad \frac{\Sigma y^2}{\sigma_y^2} = Y, \quad \frac{\Sigma xy}{\sigma_x\sigma_y} = Z,$$

we have

$$T_1\sigma_x^2 = \sigma_x^4 - \frac{\lambda_1}{2}\left[\frac{n}{2} - \frac{1}{2(1-r^2)}(X - rZ)\right]$$

$$T_2 \sigma_y^2 = \sigma_y^4 - \frac{\lambda_2}{2}\left[\frac{n}{2} - \frac{1}{2(1-r^2)}(Y - rZ)\right]$$

$$T_3 = r + \frac{\lambda_3}{2}\left[\frac{rn}{1-r^2} - \frac{r}{(1-r^2)^2}\left(X - \frac{1+r^2}{r}Z + Y\right)\right]$$

It is obvious that these equations do not take the required form and a strict simultaneous solution is not possible. If, however, we assume that T_1 and T_2 are sufficient statistics, we have, as in the maximum likelihood method, from and

$$\frac{n}{2} = \frac{1}{2(1-r^2)}(X - rZ),$$

$$\frac{n}{2} = \frac{1}{2(1-r^2)}(Y - rZ);$$

whence substituting for X and Y in terms of Z in equation we get

$$T_3 = r + \frac{\lambda_3}{2}\left[-\frac{rn}{1-r^2} + \frac{1}{1-r^2}Z\right]$$

i.e.

$$T_3 = r + \frac{\lambda_3}{2}\left[-\frac{rn}{1-r^2} + \frac{1}{1-r^2}\frac{\Sigma xy}{\sigma_x \sigma_y}\right].$$

$\frac{\lambda_3}{2} = \frac{1-r^2}{n}$ gives $T_3 = \frac{\Sigma xy}{n \sigma_x \sigma_y}$,

the corrected result being

$$\frac{\Sigma xy}{n-1}\frac{1}{\sigma_x \sigma_y},$$

3·36· <u>The Gamma Distribution</u>

From the preceding examples it is seen that the simultaneous solution for unbiassed statistics does not appear to be in very

frequent requisition. It is also seen that the simultaneous solution of the likelihood equations does not usually lead to unbiassed statistics, as in the case of the simultaneous location and scaling of the normal curve, where the method of maximum likelihood leads to the biassed statistic

$$\hat{\sigma}^2 = \frac{1}{n} \Sigma (x_r - \bar{x})^2.$$

However a solution is sometimes possible, as the following example will show.

To estimate θ_1, θ_2 and θ_3 in

$$\varphi(x|\theta_1,\theta_2,\theta_3) = \frac{1}{\theta_2(\theta_3!)} \left(\frac{x-\theta_1}{\theta_2}\right)^{\theta_3} e^{-\frac{x-\theta_1}{\theta_2}}$$

$$\log \varphi = -(1+\theta_3)\log\theta_2 - \log(\theta_3!) + \theta_3 \log(x-\theta_1) - \frac{x-\theta_1}{\theta_2}.$$

For θ_1

$F_2 = \frac{1}{\theta_2}$ but F_1 cannot be found; hence there exists no sufficient statistic for θ_1 or any function of θ_1.

We must therefore regard θ_1 as given if we wish to find sufficient statistics for θ_2 and θ_3.

In his 1921 paper Fisher solves simultaneously for θ_1, θ_2 and θ_3 by solving

$$\frac{\partial}{\partial\theta_1}\log\Phi = 0 \; ; \; \frac{\partial}{\partial\theta_2}\log\Phi = 0 \; ; \; \frac{\partial}{\partial\theta_3}\log\Phi = 0.$$

It happens that θ_2 and θ_3 do yield sufficient statistics when θ_1 is regarded as known. But it is clear that no sufficient set of statistics exists for the simultaneous estimation of θ_1, θ_2 and θ_3. Hence the method used by Fisher is rather questionable.

Regarding θ_1 as known, we may now write the curve in the form

$$\phi(x \mid \theta_2, \theta_3) = \frac{1}{\theta_2(\theta_3!)} \left(\frac{x}{\theta_2}\right)^{\theta_3} e^{-\frac{x}{\theta_2}},$$

$$\log \phi = -(1+\theta_3)\log\theta_2 - \log(\theta_3!) + \theta_3 \log x - \frac{x}{\theta_2}$$

For θ_2

$$T_1 = -\frac{1}{\theta_2} \quad ; \quad T_2 = -(1+\theta_3)\log\theta_2 \; ;$$

$$-\frac{dT_2}{dT_1} = \theta_2(1+\theta_3) = t_2 ,$$

$$\therefore \theta_2 = t_2/(1+\theta_3).$$

$$\log \Phi = -n(1+\theta_3)\log\frac{t_2}{1+\theta_3} - n\log(\theta_3!) + \theta_3(\Sigma \log x) - \frac{\Sigma x}{t_2}(1+\theta_3).$$

$$\therefore \frac{\partial}{\partial t_2}\log\Phi = -\frac{n(1+\theta_3)}{t_2} + \frac{\Sigma x}{t_2^2}(1+\theta_3)$$

$$= -\frac{n(1+\theta_3)}{t_2^2}\left\{t_2 - \frac{\Sigma x}{n}\right\}.$$

Whence

$$T_2 = t_2 + \frac{\lambda_2}{2}\frac{n(1+\theta_3)}{t_2^2}\left\{t_2 - \frac{\Sigma x}{n}\right\}.$$

$\frac{\lambda_2}{2} = \frac{t_2^2}{n(1+\theta_3)}$ gives $T_2 = \frac{\Sigma x}{n}$

i.e.

$$\hat{\theta}_2(1+\theta_3) = \frac{1}{n}\Sigma x . \qquad \text{---(1)}$$

For θ_3

$$F_2 = -\theta_3 \log \theta_2 - \log(\theta_3!),$$

$$F_1 = \theta_3 ;$$

$$-\frac{\partial F_2}{\partial F_1} = \log \theta_2 + F(\theta_3),$$

where

$$F(\theta_3) = \frac{\partial}{\partial \theta_3}\{\log(\theta_3!)\}.$$

Let us estimate for $F(\theta_3) = t_3$.

$$\frac{\partial}{\partial t_3} \log \Phi = \frac{\partial}{\partial \theta_3} \frac{\partial \theta_3}{\partial t_3} \log \Phi$$

$$= \frac{1}{F'(\theta_3)} \left\{ -n F(\theta_3) + \Sigma \log \frac{x}{\theta_2} \right\}$$

$$= \frac{2}{\lambda_3} \{ F(\theta_3) - T_3 \} ;$$

i.e.

$$T_3 = F(\theta_3) + \frac{\lambda_3}{2} \frac{1}{F'(\theta_3)} \left\{ n F(\theta_3) - \Sigma \log \frac{x}{\theta_2} \right\}.$$

$$\frac{\lambda_3}{2} = -\frac{F'(\theta_3)}{n}$$

gives

$$T_3 = \frac{1}{n} \Sigma \log \frac{x}{\theta_2}$$

$$= \frac{1}{n} \Sigma \log x - \log \theta_2 .$$

Hence

$$F(\hat{\theta}_3) + \log \theta_2 = \frac{1}{n} \Sigma \log x . \quad \text{---(2)}$$

(1) gives the minimal variance estimate for θ_2 corresponding to a fixed θ_3 ; and (2) gives the minimal variance estimate for θ_3 corresponding to a fixed θ_2 . Hence for $\hat{\theta}_2$ and $\hat{\theta}_3$ we solve simultaneously

$$\left.\begin{aligned} \hat{\theta}_2 (1 + \hat{\theta}_3) &= \frac{1}{n} \Sigma x \\ F(\hat{\theta}_3) + \log \hat{\theta}_2 &= \frac{1}{n} \Sigma \log x \end{aligned}\right\} .$$

CHAPTER FOUR.

The Empirical Methods.

4.00 The so-called "empirical" methods of estimation are based upon the known sampling distributions of certain functions of the observations composing a sample of drawn from a population of known form. The chief exponents of these methods, which are, of course, more limited in scope than maximum likelihood, are W.E.Deming R. T. Birge and E.J.G. Pitman. A short account of these methods is given in this chapter.

4.01 The Normal Curve.

Let us start with the problem of locating and scaling the normal curve when we have at our disposal a sample of n whose mean is u and standard deviation s .

The joint distribution of u and s is known to be

$$y\,du\,ds = \left[\frac{N\sqrt{n}}{\sigma\sqrt{2\pi}}\, e^{-nu^2/(2\sigma^2)}\, du\right]\left[\frac{n^{\frac{1}{2}(n-1)}}{\Gamma\left[\frac{1}{2}(n-1)\right]2^{\frac{1}{2}(n-3)}\sigma}\left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds\right] \quad \text{--- (1)}$$

Integrating this with respect to s , from 0 to ∞ , we get for the distribution of u

$$y\,du = \frac{N\sqrt{n}}{\sigma\sqrt{2\pi}}\, e^{-nu^2/(2\sigma^2)}\, du, \quad \cdots (2)$$

a normal curve.

This gives the sampling variance of the mean of a sample of n as σ^2/n.

Integrating (1) with respect to u from $-\infty$ to ∞ we get, as the sampling distribution of s,

$$y\,ds = \frac{N n^{\frac{1}{2}(n-1)}}{\Gamma[\frac{1}{2}(n-1)]\,2^{\frac{1}{2}(n-3)}\,\sigma}\left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}}\, ds, \quad \cdots (3)$$

N denotes an indefinitely large number of samples of n : σ is the standard deviation of the parentpopulation.

Deming and Birge, [Review of Modern Physics, vol.6, No.3, July 1934 ,] call equation (3) "Helmert's Equation," as it is equivalent to a result obtained by F. R. Helmert in 1876.

4·02. Maximum Likelihood Solution.

Now equation (3) represents the probability that, given σ , s should fall in the range ds . Hence Deming has argued that out of a mythical "infinity of populations" there is one that is most favourable to the occurrence of

our particular sample, and that it can be found by maximising y in equation (3) for variations in σ when s is fixed. The best value $\hat{\sigma}$ of σ is therefore that which maximises the expression of the right hand side without the differential element.

From this we find

$$\hat{\sigma} = s\left[n/(n-1)\right]^{\frac{1}{2}}.$$

In other words we get the maximum likelihood solution.

Actually there is a great difference between this method and Fisher's method of maximum likelihood. The above arguments are a concealed form of the inverse probability argument. For we have assumed that there are an infinity of populations, each with a different value of σ, hence we are tacitly assuming that σ itself has a distribution. But if nothing is known concerning σ, this appears to be an unjustified assumption. Moreover if nothing is known concerning the parameter σ which specifies the population we are surely not justified in speaking of the "probability of drawing a sample of $S.D.\ s \pm \frac{1}{2}\alpha s$." [Deming & Birge, loc. cit., p.145.] The notion of likelihood introduced by Fisher to serve, instead of probability, for arguments of the above type frees us from these contradictions.

4·03. The "Mean Estimate."

To return to our problem of estimating σ by means of equations (3)

Assume that the observed standard deviation is the mean of the various standard deviations obtained by taking a large number of samples of n from the parent population.

This is equivalent to placing s at the mean, $\bar{s}$, of Helmert's distribution.

For this we have

$$\bar{s} = \frac{\sigma (2\pi/n)^{\frac{1}{2}}}{B[\frac{1}{2}(n-1), \frac{1}{2}]} .$$

Hence

$$\hat{\sigma}_M = s(n/(2\pi))^{\frac{1}{2}} B[\tfrac{1}{2}(n-1), \tfrac{1}{2}]$$

$$\fallingdotseq s/(1 - \tfrac{3}{4n} - \tfrac{7}{32n^2} - \cdots)$$

4·04. The "Median Estimate."

Regard s as the median of the standard deviations of a large number of samples of n .

$\hat{\sigma}_{Me}$ is then given by

$$s = \hat{\sigma}_{Me}/f \quad ; \quad \text{i.e.} \quad \hat{\sigma}_{Me} = sf$$

where

$$\int_0^{\sigma/f} y\, ds = \tfrac{1}{2} \int_0^{\infty} y\, ds = \tfrac{1}{2} N,$$

y being given by Helmert's Equation.

4·05 The "Modal Estimate".

Similarly, by placing s at the mode of the standard deviations obtained from various samples of n, we get

$$\hat{\sigma}_{Mo} = s\left[n/(n-2)\right]^{\frac{1}{2}}.$$

Regard s^2 as the mean square of the $s \cdot d \cdot$'s of a large number of samples of n.

This gives

$$\hat{\sigma}_{MS} = s\left[n/(n-1)\right]^{\frac{1}{2}},$$

the maximum likelihood estimate.

4·06 The "Best" Estimate.

It is obvious that all the above estimates, (and there is no limit to the number of such estimates), are consistent statistics, and as n increases they all tend to equivalence with s, which in turn tends to σ the standard deviation of the population.

The function s of the observations is what E. J. G. Pitman calls a "sufficient estimator" for σ, but it does not of course follow that all the above statistics are all sufficient statistics. The choice of the factor which multiplies s (the factor being merely a function of n) will, in each case, determine the variance of our estimate. We have decided to choose this factor so as to make the variance of the estimate a minimum for

n finite . In this case we have the estimate

$$\hat{\sigma} = s[n/(n-1)]^{\frac{1}{2}} ;$$

and we have justified this procedure on the grounds of "reliability of estimate."

We will see later that an entirely different conception of the "best estimate," such as Pitman's conception of the "closest" estimate leads to a different expression for $\hat{\sigma}$.

4.07 Significance of the Estimates.

In order to test the relative significance of the various estimates, Deming and Birge adopt the following procedure.

Each of the above methods results in an expression for $\hat{\sigma}$ in the form

$$\hat{\sigma} = \omega s$$

where ω is a function of n which $\rightsquigarrow 1$ as $n \rightsquigarrow \infty$.

Hence

$$\hat{\sigma} - \sigma = \omega s - \sigma$$

$\therefore$ "mean square error" $= E(\hat{\sigma} - \sigma)^2$

$$= \frac{1}{N} \int_0^\infty (\omega s - \sigma)^2 y \, ds$$

$$= \frac{\sigma^2}{n} \int_0^\infty (1 - 2\omega s/\sigma + \omega^2 s^2/\sigma^2) y \, ds . \quad \text{---(1)}$$

From the zero, first, and second moments of Helmert's curve, this is found to be

$$E(\hat{\sigma}-\sigma)^2 = \sigma^2\{1 - 2\omega\bar{5}/\sigma + \omega^2(n-1)/n\},$$

where

$$\bar{5} = \sigma\sqrt{\tfrac{2\pi}{n}}\Big/ B[\tfrac{1}{2}(n-1), \tfrac{1}{2}].$$

Hence

$$\frac{\{E(\hat{\sigma}-\sigma)^2\}^{\frac{1}{2}}}{\sigma} = \{1 - 2\omega\bar{5}/\sigma + \omega^2(n-1)/n\}^{\frac{1}{2}} \quad \cdots (2)$$

As σ is unknown, and we have only $\hat{\sigma}$, they form what they call the "proportional root mean square error" (p·r·m·s· error), (it is simply K. Pearson's "coefficient of variation"),

$$\frac{\{E(\hat{\sigma}-\sigma)^2\}^{\frac{1}{2}}}{\hat{\sigma}} = \frac{\sigma}{\omega s}\{1 - 2\omega\bar{5}/\sigma + \omega^2(n-1)/n\}^{\frac{1}{2}} \quad \cdots (3)$$

The R.H.S. of (3) has its minimum value for variations in ω when $\omega = \sigma/\bar{5}$; i.e. the mean estimate (a) given above has the minimum p·r·m·s· error.

The R.H.S. of equation (2) is denoted by F, and a table is given of values of F for the "maximum likelihood" solution $\omega = \sqrt{\frac{n}{n-1}}$, and the "mean estimate," for which $\omega = \sqrt{\frac{n}{2\pi}}\, B[\frac{1}{2}(n-1), \frac{1}{2}]$.

Values of F corresponding to $n = 2$, 4 & 9 are given below; also values of F corresponding to the median estimate.

	F Max. Likelihood	F Mean estimate	F Median estimate
2	0 636	0 756	0 899
4	0 397	0 422	0 425
9	0 248	0 254	0 257

4.08. Criticism of "p.r.m.s error."

Actually, neither the "root mean square error" nor the p.r.m.s. error have gained much acceptance as tests of significance. The variance is a much more satisfactory test of significance. The "mean square error"

$$E(\hat{\sigma}-\sigma)^2$$

is not the same as the variance of $\hat{\sigma}$ unless the mean value of $\hat{\sigma}$ in samples of n is σ.

Again the "p.r.m.s. error" is not very satisfactory, for even when $E(\hat{\sigma}-\sigma)^2$ is the variance of $\hat{\sigma}$,

$$\frac{E(\hat{\sigma}-\sigma)^2}{\hat{\sigma}^2}$$

really compares the variance of $\hat{\sigma}$ over all samples with the variance within a single sample, multiplied by ω (some function of n). But whereas $E(\hat{\sigma}-\sigma)^2$ is a certain definite quantity, independent of our particular sample, $\hat{\sigma}$ itself depends upon the actual values of $x_1, \ldots, x_n$ in the sample, and, of course, varies from sample to sample.

4.09 The Method of the Median.

We have seen in the introduction that in virtue of the invariance property of the median, the so called method of the median has many points of recommendation for practical use, despite the fact that, in general, a median estimate will not be a minimal variance estimate.

The elaboration of this method is due mainly to the work of E. J. G. Pitman, [See: Proc. Camb. Phil. Soc., 33, (1937) .]

If T is an estimate of a parameter θ, T will be a "closer" estimate of θ than T' if the probability that

$$|T-\theta| < |T'-\theta|$$

is greater than $\frac{1}{2}$.

The derivation of the "closest" estimate of θ depends upon two well known theorems concerning the median of a variate X.

I. The mean absolute deviation of X about the median, M, is a minimum; i.e.

$$\Sigma |X-M| = \min.$$

II. If M is unique, and C any fixed number not equal to M, then the probability that

$$|X-M| < |X-C|$$

is greater than $\frac{1}{2}$.

If T is a function of the observations leading to a

sufficient statistic for θ, the closest estimate of θ will be derived from the median value of T in different samples of n. If this median value of T is a function $\psi(\theta)$ of θ, then the closest estimate of θ will be $\psi^{-1}(T)$; or the closest estimate of θ is obtained by equating a sufficient "estimator", T, to its median value in samples of a given size.

A third theorem used by Pitman is as follows:

III. If X_1 is a variate with median value θ, and X_2 any other variate, then X_1 is a closer estimate of θ than X_2, (i.e. $|X_1-\theta| < |X_2-\theta|$), if there exists a third variate Z, always of the same sign, (and, if continuous, not taking the value θ) such that

$$X_1 \quad \text{and} \quad Z|X_2 - X_1|$$

are independent.

4·10· Examples.

1. Variance of a normal population. The symmetric function,

$$S_2 = \sum_{r=1}^{n} (x_r - \bar{x})^2$$

is a sufficient "estimator" from which to obtain $\hat{\sigma}^2$. Now it is known that, for a normal population, S_2/σ^2 is distributed like χ^2 for $n-1$ degrees of freedom. Hence the closest estimate of S_2/σ^2 is obtained by equating

s_2/σ^2 to its median value in samples of n, viz. K_{n-1}, say.

Hence the closest estimate of σ^2 that can be derived from s_2 is

$$\hat{\sigma}^2 = \frac{s_2}{K_{n-1}}$$

If $n > 1$, a good approximation to K_n is

$$K_n = n - \frac{2}{3} + \frac{0.09}{n}.$$

If the mean is known, and taken as origin, then

$$\hat{\sigma}^2 = \frac{s}{K_n},$$

where

$$s = \sum_{r=1}^{n} x_r^2$$

11. Type III distribution.

This may be written in the form

$$\varphi = \frac{1}{\Gamma(m)} e^{-x} x^{m-1}$$

referred to the lower end-point as origin. Pitman calls this a " $\Gamma(m)$ " distribution, and demonstrates that it possesses the following properties.

(a) The distribution of $\frac{1}{2}\chi^2$ is a $\Gamma(\frac{1}{2}m)$ distribution, where $\frac{1}{2}m$ = number of degrees of freedom; hence the median of the $\Gamma(m)$ distribution is

$$\tfrac{1}{2} K_{2m} \fallingdotseq m - \tfrac{1}{3} + \frac{0.09}{4m}$$

(b) The sum of two independent variates having respectively $\Gamma(m_1)$ and $\Gamma(m_2)$ distributions has itself a $\Gamma(m_1+m_2)$ distribution. (Easily proved from the characteristic functions $(1-\alpha)^{-m_1}$ and $(1-\alpha)^{-m_2}$).

(c) If $x_1, x_2, \cdots, x_n$ is a sample of n from a gamma population, and $F(x_1, \cdots, x_n)$ a homogeneous function of zero degree (i.e. a function independent of scale), then

$$\Sigma(x_r) \quad , \quad F(x_1, \cdots, x_n)$$

are independent.

(d) If X_1 and X_2 are independent variates with $\Gamma(m_1)$ and $\Gamma(m_2)$ distributions, and if $Z = X_1/X_2$, the probability that $Z \leq K$ is

$$\frac{1}{B(m_1,m_2)} \int_0^{\frac{K}{K+1}} x^{m_1-1}(1-x)^{m_2-1}\,dx = \frac{1}{B(m_1,m_2)} \int_0^{K} \frac{y^{m_1-1}}{(1+y)^{m_1+m_2}}\,dy .$$

The curve may be written in the form

$$\varphi = \frac{1}{c^m \Gamma(m)} e^{-\frac{2x}{c}} x^{m-1} .$$

$\sum_{r=1}^{n} x_r$ and $x_1 x_2 \cdots x_n$

are sufficient estimators for m and c respectively, separately when the other is known, or simultaneously when neither is known.

(When the end-point is unknown, no sufficient statistic for locating the curve exists, except when $m = 1$.)

To estimate c, assuming m is known:

$\Sigma x_r/c$ has a $\Gamma(mn)$ distribution (by property (b))

Its median is g_{mn}, $(= \frac{1}{2} K_{2m})$. Hence

$$\hat{c} = \frac{\Sigma x_r}{g_{mn}}$$

If c' is another estimate of c, it must be homogeneous of the first degree in x. Hence $\frac{c'}{\hat{c}}$ is homogeneous of degree zero. Hence $\hat{c}$ and $c'/\hat{c}$ are independent, and so, by Theorem III above, $\hat{c}$ is a closer estimate of c than c'.

For the likelihood solution we have

$$\Phi = c^{-nm} \{\Gamma(m)\}^{-n} e^{-\Sigma x/c} (x_1 x_2 \cdots x_n)^{m-1}.$$

$$\frac{\partial}{\partial c} \log \Phi = -\frac{nm}{c} + \frac{\Sigma x}{c^2} = 0$$

Hence $$c' = \frac{\Sigma x}{nm}$$

The difference between this and the median estimate is considerable unless nm is small.

111 The exponential distribution.

$$\varphi = \frac{1}{c} e^{-\frac{x-a}{c}}, \quad x \geq a, \quad c > 0$$

$$\Phi = c^{-n} e^{-\frac{1}{c}\Sigma(x_r - a)}$$

$$= c^{-n} e^{-\frac{1}{c}\{\Sigma(x_r - x_0) + n(x_0 - a)\}}$$

where x_0 is the smallest number of the sample. x_0 and $\Sigma(x_r - x_0)$ are sufficient estimators for a and c, for when they are fixed, Φ is fixed, and $x_r \geq a$ is replaced by $x_r \geq x_0$. Hence the distribution of any other statistic is independent of a and c. Also x_0 and $\Sigma(x_r - x_0)$ are independent.

$X = n(x_0 - a)/c$ has a $\Gamma(1)$ distribution, and

$Y = \Sigma(x_r - x_0)/c$ has a $\Gamma(n-1)$ distribution.

Hence the distribution of $Z = \frac{X}{Y}$ is known.

The probability that $Z < K$ is

$$(n-1)\int_0^{\frac{K}{K+1}} (1-x)^{n-2}\,dx = 1 - \frac{1}{(1+K)^{n-1}}$$

This gives $2^{1/(n-1)} - 1$ as the median value of Z.

Hence the probability is $\frac{1}{2}$ that

$$x_0 - a \leq (2^{\frac{1}{n-1}} - 1)\Sigma(x_r - x_0)/n.$$

$$\therefore \hat{a} = x_0 - (2^{\frac{1}{n-1}} - 1)\Sigma(x_r - x_0)/n.$$

The median value of $\Sigma(x_r - x_0)/c$ is g_{n-1}, where g_n, $(= \frac{1}{2}\kappa_{2n})$, denotes the median of the $\Gamma(n)$ distribution.

Hence

$$\hat{c} = \frac{\Sigma(x_r - x_0)}{g_{n-1}} .$$

If c' is another statistic for estimating c, we have, by the same reasoning as in example II, that $\hat{c}$ is a closer estimate than c'.

IV The rectangular distribution.

Let a be the centre of the distribution and c the range; x_0 and y_0 the least and greatest numbers of the sample.

For a: The median value of $\frac{x_0 + y_0}{2}$ is a

For c: The probability that

$$\frac{y_0 - x_0}{c} \leq K$$

is

$$nK^{n-1} - (n-1)K^n$$

Hence if

$$n\lambda^{n-1} - (n-1)\lambda^n = \tfrac{1}{2},$$

$\hat{c}$ is the median value of

$$\frac{y_0 - x_0}{\lambda} .$$

If a is known, and only c has to be estimated, then T, the greater of $|x_0 - a|$ and $|y_0 - a|$, is a sufficient statistic for estimating c.

In fact

$$\hat{c} = 2^{\frac{n+1}{n}} T .$$

4·11 General Procedure.

In general it would appear that we can proceed as follows:

If we desire the closest estimate of a parameter θ, we might, by the method of maximum likelihood or otherwise find a sufficient estimator $F(x_1, \cdots, x_n)$. If then we know the sampling distribution of $\frac{F}{\theta}$ we know its median K; or, possibly we can find that median by other means. Then we have, as the closest estimate of θ,

$$\hat{\theta} = F/K .$$

4·12. Remarks on the Empirical Methods.

The moment generating function of the distribution of a function $T(x_1, \cdots, x_n)$ of the observations $x_1, \cdots, x_n$ composing a sample of n is

$$\iint \cdots \Phi(x_1, \cdots, x_n) \cdot e^{T\alpha} \, dx_1 \cdots dx_n .$$

The mean of the sampling distribution is the coefficient of α in this; i.e. is

$$\iiint \cdots \Phi \cdot T \cdot dx_1 \cdots dx_n .$$

Hence to put the "best" estimate, $\hat{\theta}$, of θ at the mean of the sampling distribution of T in samples of n is to put

$$\iiint \cdots \Phi \cdot T \cdot dx_1 \cdots dx_n = \hat{\theta}$$

The equation

$$\iiint \cdots T(x_i) \cdot \Phi(x_i|\theta)\, dx_1 \cdots dx_n = \theta$$

leads to an equation between T and θ, satisfied by a number of functions T_j of the observations. The method of consistent minimum variance chooses that particular function T which has the ~~best~~ least variance among the T_j.

The method of Deming, Pitman, etc. is to assume that we know T save for some factor (such as Deming's ω), and that the problem of estimation consists simply in the finding of ω. This is of course to assume the problem of estimation half solved.

The empirical methods place $\hat{\theta}$ at the mean, median or mode, as the case may be, of the sampling distribution of T.

4·13. The Normal Curve.

For example, it is well known that the distribution of

$$T = \frac{1}{n}\Sigma(x_r - \bar{x})^2,$$

where the sample $x_1, \cdots\cdots, x_n$ is drawn from a normal population has mean

$$\frac{n-1}{n}\sigma^2.$$

Hence our mean estimate of σ^2 is given by

$$\frac{1}{n}\Sigma(x_r - \bar{x})^2 = \frac{n-1}{n}\hat{\sigma}^2$$

i.e. $$\hat{\sigma}^2 = \frac{1}{n-1}\Sigma(x_r - \bar{x})^2$$

where $$\bar{x} = \frac{1}{n}\Sigma x_r .$$

This method of approach is not the same thing as that of estimating the "best" value of σ^2 in

$$\frac{1}{\sqrt{2\pi}\cdot\sigma}\, e^{-\frac{1}{2}(x-m)^2/\sigma^2}$$

by the postulates of consistency and minimum sampling variance.

As the problem is posed by Deming, the choice of an estimator $\Sigma(x_r - \bar{x})^2$ is an arbitrary process, unless the choice is justified on other grounds. Theoretically there are an infinite number of functions T_j of the x_r with mean value equal to the mean of a given frequency distribution, and

there is no obvious reason (though of course reasons can be given) why we should choose Σx_r, for instance, as an "estimator" for m in $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-m)^2}$, rather than any other linear, (or, for that matter, non-linear) function of the x_r, such as the harmonic mean, or the cube-root of mean cubes. It is only when we make use of a second postulate, such as that of minimum sampling variance, that the problem is fully solved, without the introduction of estimators.

4·14. Various "mean" Estimates.

The dangers that attend the unqualified presentation of the method of the mean as given by Deming and Birge, and its uselessness in that particular form may be demonstrated by the following arguments.

The distribution of

$$v_{11} = s^2 = \frac{1}{n}\Sigma(x_r - \bar{x})^2$$

in samples of n from a normal population is

$$\int dv = \frac{1}{\Gamma[\frac{1}{2}(n-1)]} \left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}} v_{11}^{\frac{n-3}{2}} e^{-\frac{n v_{11}}{2\sigma^2}} dv \quad \ldots (1)$$

If we put $v_{11} = s^2$, $dv_{11} = 2s\,ds$, we get for the distribution of s

$$f\,ds = \frac{1}{\Gamma[\frac{1}{2}(n-1)]}\left(\frac{n}{2\sigma^2}\right)^{\frac{n-1}{2}} s^{n-3}\, e^{-\frac{ns^2}{2\sigma^2}} \cdot 2s\,ds$$

$$= \frac{1}{\Gamma[\frac{1}{2}(n-1)]}\;\frac{1}{2^{\frac{1}{2}(n-3)}}\left(\frac{n}{\sigma^2}\right)^{\frac{n-1}{2}} s^{n-2}\, e^{-\frac{ns^2}{2\sigma^2}}\,ds$$

$$= \frac{n^{\frac{n-1}{2}}}{\Gamma[\frac{1}{2}(n-1)]\cdot 2^{\frac{1}{2}(n-3)}\cdot\sigma}\left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}}\,ds. \qquad \ldots (2)$$

This is equivalent to the result given by Deming and Birge as "Helmert's Equation."

Now the mean value of v_{11} in (1) is

$$\frac{n-1}{n}\sigma^2 ;$$

and this is also the mean value of s^2 in (2).

But the mean value of s in (2) is

$$\sigma\left(\frac{2}{n}\right)^{\frac{1}{2}} \Big/ B\left[\tfrac{1}{2}(n-1), \tfrac{1}{2}\right].$$

Deming and Birge say that, for the mean estimate of a parameter, we equate a sufficient estimator to its mean in samples of a given size, viz. n.

(a) The mean estimate of v_{11} is obtained by putting

$$\frac{\Sigma(x_r-\bar{x})^2}{n} = \frac{n-1}{n}\hat{\sigma}^2, \quad \text{its mean in (1)};$$

$$\text{i.e.} \quad \hat{\sigma}^2 = \frac{\Sigma(x_r-\bar{x})^2}{n-1} = \frac{s_2}{n-1}$$

(b) Or we might put

$$s^2 = \frac{\Sigma(x_r - \bar{x})^2}{n} = \frac{n-1}{n}\hat{\sigma}^2 \text{ , the mean of } s^2 \text{ in (2)},$$

and get, again,

$$\hat{\sigma}^2 = \frac{s_2^2}{n-1}.$$

(c) But, according to Deming and Birge, we might equally well put s, $\left[= \left\{\frac{1}{n}\Sigma(x_r - \bar{x})^2\right\}^{\frac{1}{2}}\right]$, equal to the mean of s in (2), and get

$$\frac{1}{n}\Sigma(x_r - \bar{x})^2 = \hat{\sigma}^2\left(\frac{2\pi}{n}\right) \Big/ \left\{B\left[\tfrac{1}{2}(n-1), \tfrac{1}{2}\right]\right\}^2$$

i.e.
$$\hat{\sigma}^2 = s_2^2\left(\frac{n}{2\pi}\right)\left\{B\left[\tfrac{1}{2}(n-1), \tfrac{1}{2}\right]\right\}^2,$$

which is inconsistent with (a) and (b)

4·15 "Estimators" and "Statistics".

There is, of course, no end to the number of estimates of σ that we might obtain in this way. We might put $\hat{\sigma}^3$ equal to the mean value, in samples of n, of s^3, and so on. All such estimates, while giving different values for $\hat{\sigma}$ will be consistent in the sense that they all tend to equivalence with σ as the sample tends in size to the population. But whilst s_2 is a sufficient estimator for σ^2, it cannot be the that all statistics obtained in the above manner will be

sufficient statistics. When we estimate for θ we must at the same time be estimating for all functions of θ. Hence it is obvious that the method of the mean estimate as stated above is unsatisfactory as it does not lead in that form to the "best" estimate. As we have seen, we get the best estimate only when the condition of the mean, i.e. the condition that our estimate should be unbiassed, is combined with a second condition, such as that of minimum variance. In such a case we find the "basic function" of θ, i.e. the function of θ for which we must estimate to obtain a minimal variance statistic. In the case of the variance of a normal curve we have seen that such a function is σ^2. If then the method of the mean is applied to σ^2 we get the best estimate of σ in the form

$$\hat{\sigma} = \left[n/(n-1)\, s_2 \right]^{\frac{1}{2}}$$

The difficulties outlined above arise, of course, from the fact that the mean of squares, for example, is not the same thing as the square of the mean of any number of values x_r.

4.16. Method of the Median.

But the method of the median is not open to the same objection, for it is the case that the median of squares is the same as the square of the median, and so on. Accordingly it makes no difference whether we apply method of the median

to estimating θ or θ^2 or any function of θ that is monotone within the range of possible values of θ . Moreover the method of the median saves us from the task of finding the particular function of θ for which we must estimate.

4·17 The "Posterior" Method.

This method is based upon Laplace's generalisation of Bayes' Theorem, and depends upon the notion of inverse probability, the validity of which has been strongly contested, particularly, in recent years by R. A. Fisher in a series of papers written between 1930 and 1934.

It will suffice here to summarise the method, which, in brief, consists in assuming that the population for which we desire to find θ is itself one member of a population of populations each of which possess a different value of θ . In other words it assumes a prior distribution of θ whose form depends upon our prior knowledge (or lack of knowledge) concerning θ , or upon assumptions that we may make concerning θ . By combining this prior "knowledge" with the actual knowledge furnished by the sample we may, in the following manner, obtain a new distribution which will serve to define more precisely the range in which θ should lie.

4·18 Normal Curve.

If $\varphi(\alpha)$ is, in the case of the normal curve for example,

the ordinate on the prior existence curve at the abscissa σ, then the probability that σ lies in the range $d\sigma$ is

$$\varphi(\sigma)\, d\sigma. \qquad \text{---} \quad (1)$$

This is called the "prior existence probability."

Now, by Helmert's Equation, the probability of drawing a sample with standard deviation $s \pm \frac{1}{2} ds$ from a population with standard deviation σ is

$$\text{constant} \times \sigma^{-1} \left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}} ds. \qquad \text{---} \quad (2)$$

This is called the "prior productive probability" of σ.

Hence the probability of drawing our sample, subject to the condition is

$$p\, d\sigma\, ds = \text{const.} \times \sigma^{-1} \left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}} \cdot \varphi(\sigma)\, d\sigma\, ds, \qquad \text{---} \quad (3)$$

this determining a surface for the joint distribution of s and σ.

Hence the probability that the standard deviation of the population lies in the range

$$\sigma \pm \tfrac{1}{2} d\sigma,$$

when once the sample has yielded standard deviation s is

$$p\, d\sigma = \text{const} \times \sigma^{-1} \left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}} \varphi(\sigma)\, d\sigma. \qquad \text{---} \quad (4)$$

This is called the "posterior probability" of σ. If we choose the constant so that the total probability over all

samples is unity, we get

$$p\,d\sigma = \frac{\varphi(\sigma)\,\sigma^{-1}\left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}}}{\int_0^\infty \varphi(\sigma)\,\sigma^{-1}\left(\frac{s}{\sigma}\right)^{n-2} e^{-\frac{ns^2}{2\sigma^2}}\,d\sigma}\,d\sigma \quad \cdots (5)$$

The following example is given by Deming and Birge, (loc. cit.)

The prior existence curve is a rectangle given by

$$\varphi(\sigma) = 0 \quad , \quad 0 < \sigma < 1 \; ;$$

$$\varphi(\sigma) = 1 \quad , \quad 1 < \sigma < 2 \; ;$$

$$\varphi(\sigma) = 0 \quad , \quad 2 < \sigma .$$

Further we are given

$$n = 6 \quad , \quad s = 1.5 .$$

The posterior curve is found by substitution in equation (5), and is

$$p(\sigma) = 0 \quad , \quad 0 < \sigma < 1 \; ;$$

$$p(\sigma) = 187.13\,\sigma^{-5} e^{-\frac{27}{4\sigma^2}}\,d\sigma \quad , \quad 1 < \sigma < 2 \; ;$$

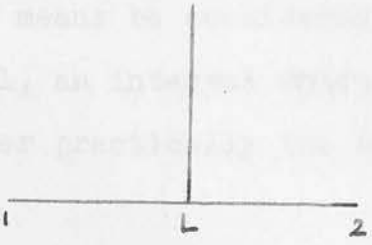
$$p(\sigma) = 0 \quad , \quad 2 < \sigma .$$

Instead of the rectangular prior distribution we now have a curve with a maximum, and having about half the area included between the ordinates at 1.46 and 1.86. In other words the posterior probability that

$1{\cdot}46 < \sigma < 1{\cdot}86$

is equal to about one-half.

If n is increased, the posterior curve becomes steeper and the probability that σ lies within the limits given above becomes greater; or the 50-50 probability interval becomes smaller. In fact, as n tends to the population, the curve will simply tend to the form



L will be the maximum likelihood solution.

4·19 Criticism of Method.

It must be noted, however, that although the probability that σ lies in a small interval covering L is now very large, there is nothing to preclude the possibility of σ lying anywhere in the interval (1, 2). In other words, even in the extreme case, σ is still not precisely defined.

Again it must be noted that the expedient of increasing n to obtain a more precise location of σ is not always a practical one. In fact the main object of any theory of estimation must be to make the greatest possible use of the actual sample with which we are provided.

From this point of view, whereas the statement that the

probability that

$$1{\cdot}46 < \sigma < 1.86$$

is 0·95 could be regarded as quite a satisfactory solution of the problem of estimation, such a statement as that the probability that

$$1{\cdot}46 < \sigma < 1{\cdot}86$$

is $\frac{1}{2}$, can by no means be considered satisfactory. In fact, when n is small, an interval which contained σ in 95% of the cases would cover practically the whole range of possible values (1, 2).

If we took, as our prior existence curve a rectangle defined by

$$\varphi(\sigma) = 0 \quad , \quad 0 < \sigma < 1{\cdot}1 \; ;$$

$$\varphi(\sigma) = 1 \quad , \quad 1{\cdot}1 < \sigma < 1{\cdot}9 \; ;$$

$$\varphi(\sigma) = 0 \quad , \quad 1{\cdot}9 < \sigma \; ;$$

we would get a different value for the probability that

$$1{\cdot}46 < \sigma < 1{\cdot}86;$$

in which case the arbitrary nature of the method becomes fully apparent.

4·20. The Precision of an Estimate.

The following device might be suggested in order to overcome the difficulty when n is fixed (= 6) and there is no possibility of increasing n :

Having found our first posterior probability curve to be

$$p\,d\sigma = 187{\cdot}13\,\sigma^{-5}e^{-\frac{27}{4\sigma^2}}d\sigma \quad , \quad 1 < \sigma < 2,$$

$$p\,d\sigma = 0 \quad , \text{ otherwise,}$$

we might take this as a new "prior existence" curve, and, repeating the process, find a second posterior curve. Apart from the great mathematical difficulties introduced in this way, the following general criticism will apply:

The more precise we desire to make the information concerning the value of σ (either by increasing n , or by repeating the process one or more times), the more we reject the "information" contained in the prior existence curve, and the more we rely upon the information contained in the sample itself, till finally we do not use the former at all. Hence we conclude that, in the end, the postulating of a prior existence curve is quite immaterial and also quite arbitrary, as we might postulate any kind of prior existence curve we please, and then proceed to reject the information contained in it, in the light of the information furnished by the sample

itself.

In their paper, Deming and Birge state: "It is therefore correct to say that a value of σ can be established by taking a long series of measurements." But they do not see that, in so doing, they are relegating their prior existence curve into the dim background.

4·21. Other Prior Existence Curves.

Other prior existence curves have been given by Molina and Wilkinson, (Bell Syst. Tech. J. - 8, 632-645), for σ and for μ, the mean. They are

$$\varphi(\sigma)\,d\sigma = \frac{1}{2^{\frac{c}{2}}\,\Gamma\left(\frac{c}{2}+1\right)\sigma}\left(\frac{a}{\sigma}\right)^{c+2} e^{-\frac{a^2}{2\sigma^2}}\,d\sigma,$$

$$\theta(\mu)\,d\mu = A\left[1 + \frac{1}{1+\frac{a^2}{ns^2}}\left(\frac{\bar{x}-\mu}{s}\right)^2\right]^{-\frac{1}{2}b}\,d\mu.$$

These curves allow for the expression of any particular degree of concentration of values of σ and μ that may occur in our prior knowledge. The "posterior probability" of μ and σ together is found as before, and the posterior curve for μ deduced therefrom. The use of these curves is developed in pp.154-158 of Deming and Birge's paper.

4·22. The Probability Interval.

The final point of criticism is as follows:

Such a statement as that "the probability that

$$1.46 < \sigma < 1.86$$

is about one-half" can only be taken to mean that if we took a large number of samples of 6 from the same population and applied the rules in each case, we would find that approximately 50% of the samples yielded a value of σ lying between 1·46 and 1·86, the remainder lying outside that interval, but still within the interval (1, 2). But actually there is nothing in the method that guarantees that such will actually be the case. In fact it is very unlikely that such a state of affairs will come about. After all the population value, σ, is usually something quite definite, and is not a variable at all. Hence it either does lie in the interval (1·46, 1·86) or it does not. In other words, the probability that

$$1.46 < \sigma < 1.86$$

is either unity or zero.

Such statements as the 50% probability statement given above are merely a result of the unsupported assumption that σ is a variable and as such has a distribution.

Actually, the above probability statement must be taken to imply that we select first of all from our super-population a population; then we draw from it a sample of 6 . If the

standard deviation of the sample is exactly 1.5 we retain that population. If not, we reject it. If we repeat this process a large number of times, then we will find that of the populations the samples of which we retain approximately one-half will have standard deviations lying between 1.46 and 1.86.

Whilst this statement is correct, it seems only to accentuate the uselessness of this process as a method of estimation, since we have usually, in practice to deal with a situation where each sample drawn from the population yields a <u>different</u> value for S .

CHAPTER FIVE.

Estimation by Interval.

5·00 In this chapter we shall take up the problem of estimation from a different viewpoint altogether. We have already seen that it is not sufficient merely to find a certain value T for a parameter θ, and leave it at that. But in each case we find a standard error σ_T and write our estimate in the form

$$T \pm \sigma_T .$$

If we know the distribution of T then we can write down the probability that T will lie in the interval $(-\sigma , \sigma)$. Again it is possible, knowing the distribution of T to find two values, a and b, such that the probability that T will lie within the interval (a, b) has a definite value, say $1-\epsilon$.

Such probability statements concerning T are perfectly objective and justifiable. But the question arises: what probability statements of this nature concerning θ are we justified in making. We have already seen, in the previous chapter, the difficulties attending such statements as that the probability that

$$a < \theta < b = 1-\epsilon,$$

when this statement is based upon the assumption that itself has a distribution.

Two answers to the above question have been given, neither of which is based upon the notion of inverse probability. The first is that, in certain instances, probability statements of a special kind can be made concerning θ. Such statements are the statements of fiducial probability, a notion introduced by R. A. Fisher, [Proc. Camb. Phil. Soc., 26, (1930)]. The second is that such statements as the above are only justified when they are interpreted to mean that we are finding an interval (a, b) that will contain the unknown fixed value of θ with a given relative frequency, and that we are not justified in regarding θ as a variable for any purposes whatsoever. It will not be surprising to find that these two fundamentally different concepts lead in many instances to a similarity in the mathematical form of the results. For we have already seen this happens in the two different approaches to the problem of least squares, and again in the parallel results of the method of maximum likelihood and the method of unbiassed minimal variance.

5·01 Fiducial Probability.

As we have seen, it often happens that the distribution

of a statistic T used to estimate a parameter θ can be expressed solely in terms of θ. If this is the case, then we can express the probability P that, when θ is given, T should be less than any given value, in the form

$$P = F(T|\theta).$$

If P is given a definite value, $1-\epsilon$, then in ϵ % of samples T will exceed the $(1-\epsilon)$% value corresponding to the actual value of θ.

If, on the other hand, T is given, then the value of θ satisfying

$$F(T|\theta) = 1-\epsilon$$

is called the "fiducial ϵ % value of θ" for the given value of T. If that value is θ_1, then the true value of θ will be less than θ_1 in (ϵ) % of the trials.

Having arrived at this stage we could find another value of θ, namely θ_2, where $\theta_1 < \theta_2$ and consider the fiducial probability that, when T is given, θ should be less than θ_2 but, at the same time, greater than θ_1. In other words, we now have the fiducial probability that when T is given, θ should lie between θ_1 and θ_2, this probability being, say, $1-\epsilon$. Corresponding to any particular value of ϵ we can find an interval (θ_1, θ_2).

The fiducial probability that θ should be less than θ_2

when T is given is equal to the actual probability that T should be greater than T_2 since $(1-\epsilon)+\epsilon = 1$, where θ_2 and T_2 are corresponding values in the expression

$$P = F(T|\theta).$$

This first probability can be written in the form

$$\int_{-\infty}^{\theta_2} p(\theta|T_2)\,d\theta, \quad p \equiv \text{fiducial probability of } \theta.$$

and the second in the form

$$\int_{T_2}^{\infty} p(T|\theta_2)\,dT.$$

Hence the fiducial distribution of θ, when T is given, may be written in the form

$$df = -\frac{\partial}{\partial\theta} F(T|\theta)\,d\theta;$$

while the distribution of T when θ is given is written in the form

$$df = \frac{\partial}{\partial T} F(T|\theta)\,dT.$$

Actually it would seem that the hypothesis that T is to be a sufficient statistic is immaterial to the above discussion, but it is justified in the interests of safety. It would be dangerous to make precise probability statements concerning the

population value, θ, if we were to base these statements upon non-sufficient statistics which utilise only a fraction of the information supplied by the sample.

5-02. More than one Parameter.

The extension of the concept of fiducial probability to more than one parameter is not just a matter of course. For if T is a vector of sufficient statistics for the estimation of the vector of parameters, θ, the fact that we can write down expressions of the form

$$P_i = F_i(T_i \mid \theta_i)$$

for each of the pairs (T_i , θ_i) when all the other pairs (T_j , θ_j), $j \neq i$, are considered as known, does not mean in general that we can find an expression of the form

$$P = F(T \mid \theta)$$

where T and θ denote the complete vectors. Or, on the other hand, if an expression of the form

$$P = F(T \mid \theta)$$

can be found, there is no particular reason to believe that when we get rid of all the θ_j in the joint fiducial distribution, by integrating then out, we will be left with a fiducial distribution of θ_i that is consistent with

$$P_i = F_i(T_i/\theta_i).$$

5·03. Examples.

1. As an example of the use of fiducial probability consider the distribution of

$$t = \frac{(\bar{x}-\mu)\sqrt{n}}{s},$$

where

$$\bar{x} = \frac{1}{n}\Sigma(x),$$

$$s^2 = \frac{1}{n-1}\Sigma(x-\bar{x})^2.$$

This distribution depends only on n.

Hence for any given value of n, we can calculate what value, t_1, of t will be exceeded say 5 times in 100 trials. In this case, $P = .95$, i.e. $t > t_1$, in 5 trials out of 100. Hence

$$\mu < \bar{x} - st_1/n$$

in 5 trials out of 100.

As t_1 varies we may therefore state the probability that μ is less than any assigned value, or that it lies between any assigned values; i.e. we know the fiducial probability distribution of μ.

11. If we have a sample of n from a normal population it is known that

$$t = \sqrt{\frac{n}{n+1}}\ \frac{x-\bar{x}}{s}$$

is distributed in "Student's" distribution for $n-1$ degrees of freedom.

Hence for a given t_1,

$$P\{t > t_1 | t_1\} = P\{x > \bar{x} + s t_1 \sqrt{\tfrac{n+1}{n}} \mid t_1\},$$

where the L.H.S. means "the probability that t is greater than t_1 when t_1 is given."

Hence the fiducial distribution of an observation of not yet made, is "Student's" distribution with the factor $\sqrt{1+\frac{1}{n}}$.

III. In his paper, "The Fiducial Argument in Statistical Inference," Annals of Eugenics, V1, Part 1V, (1935), Fisher gives an example of a case where the extension of the notion of fiducial probability to more than one parameter is justified.

If we have a sample of n from a normal population, $\bar{x}$ and s being the sample mean and standard deviation, then we can find the joint distribution of the mean $\bar{x}'$ and standard deviation $\bar{s}'$ of a further n' observations.

By letting $n' \to \infty$, we find from this the fiducial distribution of μ and σ in the form

$$dP = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\, e^{-\frac{n}{2\sigma^2}(\mu-\bar{x})^2}\, d\mu \cdot \frac{1}{\frac{1}{2}(n-3)!}\left\{\frac{(n-1)s^2}{2\sigma^2}\right\}^{\frac{1}{2}(n-1)} e^{-\frac{(n-1)s^2}{2\sigma^2}}\, \frac{2\,d\sigma}{\sigma}.$$

Integrating out σ, we find, for the fiducial distribution of μ,

$$df = \frac{\frac{1}{2}(n-2)!}{\frac{1}{2}(n-3)!\sqrt{\pi(n-1)}} \frac{1}{\left\{1 + \frac{n(\mu-\bar{x})^2}{(n-1)s^2}\right\}^{\frac{1}{2}n}} \frac{\sqrt{n}\, d\mu}{s}.$$

Hence we get, correctly, that $\frac{(\mu-\bar{x})\sqrt{n}}{s}$ is distributed as is t for $(n-1)$ degrees of freedom.

Similarly we get for σ

$$df = \frac{1}{\frac{1}{2}(n-3)!}\left\{\frac{(n-1)s^2}{2\sigma^2}\right\}^{\frac{1}{2}(n-1)} e^{-\frac{(n-1)s^2}{2\sigma^2}} \frac{2\,d\sigma}{\sigma},$$

$\frac{(n-1)s^2}{\sigma^2}$ being distributed as is χ^2 for $n-1$ degrees of freedom.

5·04 Locating and Scaling a Frequency Curve.

The following method is due to E. J. G. Pitman, [Biometrika, xxx, Pts. 111 and 1V, Jan.1939.]

The problem of locating and scaling a frequency curve of the form $\frac{1}{c} f\{(x-a)/c\}$ is the problem of estimating a and c.

5·05 Estimation of a.

To demonstrate the method, we give it for the case of estimating a when c is known.

Regard the sample $x_1, \cdots, x_n$ as a fixed point in an n-dimensional space, and let $(\xi_1, \cdots; \xi_n)$ be a variable point.

Let H be a function of the ξ, and let its expectation

$$E(H) = \int_w F H \, d\xi_1 \cdots d\xi_n ,$$

where

$$F = \prod_{r=1}^{n} f(\xi_r - a).$$

Let F be non-zero in the region W_+.

Put

$$\xi_1 = z_1$$

$$\xi_r = z_r + z_1 \quad , \quad r = 2, 3, \cdots, n.$$

Then

$$\int F H \, d\xi_1 \cdots d\xi_n = \int F H \, dz_1 \cdots dz_n .$$

The locus $z_1, z_2, \cdots, z_n$ all constant is a straight line parallel to

$$\xi_1 = \xi_2 = \cdots = \xi_n .$$

Let L be any line for which

$$\int_{-\infty}^{\infty} F \, dz_1 > 0$$

The family of lines L is the same for all values of a.

If $(x_1, \cdots, x_n)$ falls on some L, it is called an "observable point."

The mean value of H on L is

$$E_L(H) = \int_{-\infty}^{\infty} F H \, dz_1 \Big/ \int_{-\infty}^{\infty} F \, dz_1 .$$

If this is a constant, h, for every L, then it follows that

$$E(H) = h;$$

and if $E_L(H) > h$ for every L, then

$$E(H) > h.$$

If I' is a set of intervals on L, let

$$\int_{I'} F\,dz_1 \Big/ \int_{-\infty}^{\infty} F\,dz_1 = P\{I' \mid L\}.$$

Let ω' ≡ region formed by all the I' on every L;
and $P(\omega')$ = probability that the sample point will fall in ω'

If $P\{I' \mid L\} = \text{const.} = \alpha$ for every L,
then $P(\omega') = \alpha.$

If $P\{I' \mid L\} > \beta$ for every L, then

$$P(\omega') > \beta.$$

Let $\xi_r - a = x_r - t$.

ξ_r is a variable; a is a constant.
x is fixed ; t is variable.

Then

$$\int_{-\infty}^{\infty} F\,dz_1 = \int_{-\infty}^{\infty} f(x_1 - t) \cdots f(x_n - t)\,dt$$

$$= \int_{-\infty}^{\infty} \Phi_{(t)}\, dt \quad , \text{ say.}$$

Also

$$E_L(H) \int_{-\infty}^{\infty} \Phi_{(t)}\, dt = \int_{-\infty}^{\infty} H \cdot \Phi_{(t)}\, dt.$$

If I denote a set of non-overlapping intervals in $(-\infty, \infty)$, then the points of L corresponding to values of t lying in I will form a set of intervals I' on L.

I is termed "proper" if its end-points $A_1, A_2, \cdots$ are functions of $x_1, \cdots, x_n$ not involving a, such that I' is independent of $(x_1, \cdots, x_n)$ on L.

For this we must have

$$A_r(x_1+\lambda, \cdots, x_n+\lambda) \equiv A_r(x_1, \cdots, x_n) + \lambda .$$

We see that I depends on $x_1, \cdots, x_n$ but not on a; whilst I' depends on a but not on $x_1, \cdots, x_n$.

Change of a from a_1 to a_2 involves translation of I' along L a distance of $(a_2 - a_1)\sqrt{n}$ in the positive direction.

We have now

$$P\{I'|L\} = \int_{I'} F\,dz_i \Big/ \int_{-\infty}^{\infty} F\,dz_i$$

$$= \int_I \Phi(t)\,dt \Big/ \int_{-\infty}^{\infty} \Phi(t)\,dt.$$

Set this equal to a constant α ; i.e.

$$\int_I \Phi(t)\,dt = \alpha \int_{-\infty}^{\infty} \Phi(t)\,dt \quad \cdots (A)$$

The problem is to find I , when α is given some specific value between 0 and 1 , say .95. We could find I from equation (A), using the condition that the sum of the lengths of the intervals in I is to be a minimum.

The points of I' are called "points of acceptance." If $t = a$, then $(x_1, \cdots, x_n)$ is a point of acceptance provided a lies in $I(x_1, \cdots, x_n)$,

i.e. $$a \in I(x_1, \cdots, x_n)$$

Points of acceptance on every L form a "region of acceptance," $\omega'(a)$. The remainder of the sample space is called the "critical region," $\omega(a)$.

If α has the same value on every L , then the probability $P\{\omega'(a)\}$ that the sample point will fall in the region of acceptance is α . If $\alpha > \beta$ this

probability $> \beta$.

Change of $\underline{a}$ from a_1 to a_2 means a translation of the region of acceptance in the positive direction of the lines L if $a_2 > a_1$.

Hence if I is chosen so that the sum of its lengths, (and consequently those of I'), is a minimum for the corresponding α , then when $a = a_1$, the probability that the sample point $\underset{\Lambda}{E}$ falls within $\omega'(a_1)$ is greater than the probability that it falls in $\omega'(a_2)$.

We have

$$P\{E \in \omega'(a_1) | a_1\} > P\{E \in \omega'(a_2) | a_1\},$$

$$\therefore P\{E \in \omega(a_1) | a_1\} < P\{E \in \omega(a_2) | a_1\}.$$

But

$$P\{E \in \omega(a_2) | a_2\} = P\{E \in \omega(a_1) | a_1\},$$

$$\therefore P\{E \in \omega(a_2) | a_2\} < P\{E \in \omega(a_2) | a_1\}.$$

Hence the critical region $\omega(a)$ is "unbiassed." If the shortest I is not always unique, we replace the sign $<$ in the last statement by $\leq$.

From equation (A) we have

$$K \int_I f(x_1-a) \cdots f(x_n-a)\, da = \alpha = K \int_I \Phi(a)\, da,$$

say, where

$$\frac{1}{K} = \int_{-\infty}^{\infty} \Phi(a)\, da .$$

If $(x_1, \cdots, x_n)$ is an observable point,

$$\int_{-\infty}^{\infty} \Phi(a)\, da > 0 .$$

When α is constant, the probability that $(x_1, \cdots, x_n)$ is a point of acceptance is α ;

$$\therefore P\{a \in I(x_1, \cdots, x_n)\} = \alpha$$

If $\alpha > \beta$,

$$P\{a \in I(x_1, \cdots, x_n)\} > \beta .$$

$$K f(x_1 - a) \cdots f(x_n - a) = K \Phi(a)$$

is the "fiducial probability function for the estimation of a .

The fiducial probability that $a \in I(x_1, \cdots, x_n)$ is α .

If we set α at 0.95, say, we can then define I , and thus make a definite statement about a whenever a set of values $x_1, \cdots, x_n$ is known. In the long run we will be correct 95 times out of 100.

5·06 Example.

As an example, let us consider the location of the trapezoidal distribution

$$\left.\begin{aligned} f(x) &= 2x \quad , \quad 0 \le x \le 1 \, , \\ &= 0 \quad , \quad x < 0 \, , \, x > 1 \end{aligned}\right\}$$

or

$$\left.\begin{aligned} f(x) &= 2(x-a) \quad , \quad a \le x \le a+1 \\ &= 0 \quad , \text{ otherwise} \end{aligned}\right\}$$

Let x_s , x_L denote the smallest and largest observations in a sample of n observations,

Here

$$K \Phi_{(a)} = 2^n / (x_1 - a) \cdots (x_n - a)$$

provided

$$a \le x_s \quad , \quad x_L \le a+1 \, ,$$

i.e. $$x_L - 1 \le a \le x_s \, .$$

$\Phi_{(a)} = 0$ outside this range.

Now

$$\int_{-\infty}^{\infty} \Phi_{(a)} \, da = \int_{x_L - 1}^{x_s} \Phi_{(a)} \, da \, .$$

Since the fiducial function is monotonic and decreasing in the above interval, it follows that the shortest I will consist of a single interval with its lower end at $x_L - 1$; i.e. is $(x_L - 1 , h)$, where

$$\int_{x_L-1}^{h} \Phi(a)\, da = \alpha \int_{x_L-1}^{x_S} \Phi(a)\, da ,$$

i.e.

$$G(h) - G(x_L - 1) = \alpha \{ G(x_S) - G(x_L - 1) ,$$

where

$$G(a) = \int_0^a \Phi(a)\, da .$$

The equation for h is therefore of degree $n + 1$ in h, and has a single root in the range $(x_L - 1 , x_S)$.

The fiducial probability that

$$x_L - 1 \leq a \leq h$$

is equal to α .

5·04 Restrictions of Method.

If $K\Phi(a)$, the fiducial function, is strictly monotonic when a lies within the interval (b_1, b_2), and is zero outside that interval, the shortest I is unique, and consists of a single interval. This also holds if $K\Phi(a)$ is strictly

increasing in (b_1, b_2), strictly decreasing in (b_2, b_3) and zero outside (b_1, b_3).

It is obvious that, since any point on L lying in I has a greater probability than any point on L outside I, I must include the maximum likelihood solution, which makes $\Phi_{(a)}$ a maximum.

$K\Phi_{(a)}$ will conform to the above conditions if $f(x)$ is strictly monotonic over a certain range of x, and zero outside that range; or if $\log f(x)$ is a concave function of x over a certain range, and zero outside that range.

One of these two conditions is satisfied by each of the following distributions: normal, gamma, beta (except when -shaped), triangular, trapezoidal (except when rectangular).

5·08 Estimation of C.

The estimation of C where the probability function of X is

$$c^{-1} f\left(\frac{x}{c}\right), \quad c > 0,$$

follows similar lines if we make the transformation

$$\log c = \gamma.$$

The distribution of $\log X$ is

$$e^{x-\gamma} f(e^{x-\gamma}),$$

γ being analogous to a in the previous case.

If $C(x_1, \cdots, x_n)$ is an estimator for C, it must be non-negative, and such that

$$C(\lambda x_1, \cdots, \lambda x_n) = \lambda C(x_1; \cdots; x_n) \quad , \quad \lambda \geq 0;$$

i.e. it must be a positive homogeneous function of the first degree in x.

Its logarithm, G, is a γ-estimator, and is such that

$$G(\lambda x_1; \cdots \lambda x_n) = G(x_1, \cdots, x_n) + \log \lambda .$$

A half line, or say, with one end at the origin is denoted by R if it belongs to W_t.

Any point on R is "observable."

The mean of H on R is

$$E_R(H) = \int_0^\infty F H r^{n-1} dr \Big/ \int_0^\infty F r^{n-1} dr,$$

where $r = \sqrt{\Sigma(\xi_r^2)}$, the distance of $(\xi_1, \cdots, \xi_n)$ from the origin.

If $E_R(H)$ has the same value h on every R, $E(H) = h$. If $E_R(H) > h$ on every R, $E(H) > h$.

For a set of intervals I' on R, $P\{I'|R\}$ is such that

$$P\{I'|R\}\int_0^\infty F r^{n-1}\,dr = \int_{I'} F r^{n-1}\,dr.$$

As before, if ω' is the region generated by the I', then

$P(\omega') = \alpha$ if $P\{I'|R\} = \alpha$ for every R ; and

$P(\omega') > \beta$ if $P\{I'|R\} > \beta$ for every R ; where $P(\omega')$ is the probability that the sample point falls in ω'.

As before we then put

$$e^{-\gamma}\xi_r = e^{-t}x_r \quad , \quad r = 1, 2, \cdots, n ;$$

where ξ is variable and γ fixed,

and x is fixed and t variable.

For points on R

$$F = e^{-nt} f(e^{-t}x_1)\cdots f(e^{-t}x_n)$$

$$E_R(H) = \frac{\int_0^\infty H e^{-nt}\cdot f(e^{-t}x_1)\cdots f(e^{-t}x_n)\,dt}{\int_0^\infty e^{-nt}\cdot f(e^{-t}x_1)\cdots f(e^{-t}x_n)\,dt}$$

$$= \int_0^\infty H\,\Phi(t)\,dt \Big/ \int_0^\infty \Phi(t)\,dt \quad , \text{ say.}$$

If I is a set of non-overlapping intervals in $(-\infty, \infty)$, the points on R corresponding to values of t lying in I, are called "points of acceptance" and correspond to a set of intervals I' on L.

If I' is independent of $(x_1, \cdots, x_n)$, I is "proper". For this, the end-points of I must be γ-estimators.

If

$$\alpha = P\{I' / R\}$$

$$\int_I \Phi(t)\, dt = \alpha \int_{-\infty}^{\infty} \Phi(t)\, dt.$$

Points of acceptance on all the R form a region of acceptance $\omega'(\gamma)$. It can be shown that the critical region obtained by using on every R the shortest I for the corresponding α is unbiassed.

For an observable point

$$\int_{-\infty}^{\infty} \Phi(t)\, dt \neq 0.$$

As before, the fiducial function for the estimation of is

$$K\Phi(\gamma) = Ke^{-n\gamma} f(x_1 e^{-\gamma}) \cdots f(x_n e^{-\gamma}).$$

If

$$\gamma \in I(x_1, \cdots, x_n)$$

is equivalent to

$$c \in J(x_1, \cdots, x_n),$$

the end-points of J are c-estimators; and if

$$K \int_I \Phi(\gamma)\, d\gamma = K \int_J c^{-n-1} f\left(\frac{x_1}{c}\right) \cdots f\left(\frac{x_n}{c}\right) dc$$

$$= K \int_J \Phi(c)\, dc \quad , \text{ say,}$$

then J is "proper" for the estimation of c. $K\Phi(c)$ is the fiducial function for the estimation of c.

5·09 Scaling the Normal Curve.

As an example, consider the scaling of the normal curve

$$\frac{1}{c\sqrt{2\pi}} e^{-\frac{1}{2}x^2/c^2}$$

Here

$$K\Phi(c) = K c^{-n-1} e^{-\frac{1}{2}s/c^2},$$

$$s = \sum_{r=1}^{n} (x_r^2).$$

$C = \sqrt{\frac{1}{2}S/h}$ is an estimator, where h is any positive homogeneous function of degree 0 in the x_r.

Hence C is an end-point of the interval I.

Determine h so that

$$\int_C^\infty K \Phi(c)\, dc = \alpha \quad (\text{const.}) \qquad \ldots \quad (1)$$

Then

$$P\{c \geq C\} = \alpha ,$$

i.e.

$$P\{\tfrac{1}{2}S/c^2 \leq h\} = \alpha . \qquad \ldots \quad (2)$$

Putting $\frac{1}{2}S/c^2 = u$ in (1) we get

$$\frac{1}{\Gamma(\frac{1}{2}n)} \int_0^h e^{-u} u^{\frac{1}{2}n-1}\, du = \alpha \qquad \ldots \quad (3)$$

Hence h is constant.

If h is any given positive number, and α is determined by (3), then (2) is true; i.e. for this curve the variate $\frac{1}{2}S/c^2$ has a $\Gamma(\frac{1}{2}n)$ distribution.

If

$$\frac{1}{\Gamma(\frac{1}{2}n)} \int_{h_1}^{h_2} e^{-u} u^{\frac{1}{2}n-1}\, du = \alpha , \qquad \ldots \quad (4)$$

then

$$P\{h_1 \leq \tfrac{1}{2}S/c^2 \leq h_2\} = \alpha \ ;$$

$$\therefore\ P\{\tfrac{1}{2}\log(\tfrac{1}{2}S/h_2) \leq \gamma \leq \tfrac{1}{2}\log(\tfrac{1}{2}S/h_1)\} = \alpha \ .$$

Hence fiducial ranges for c^2 and γ can be determined for any given value of α.

For the shortest range for γ,

$$\tfrac{1}{2}\log(\tfrac{1}{2}S/h_1) - \tfrac{1}{2}\log(\tfrac{1}{2}S/h_2) = \text{minimum}$$

$$\therefore \log h_2 - \log h_1 = \text{minimum}$$

$$\therefore \frac{dh_2}{h_2} = \frac{dh_1}{h_1} \ .$$

But from (4)

$$e^{-h_2} h_2^{\frac{1}{2}n-1}\, dh_2 = e^{-h_1} h_1^{\frac{1}{2}n-1}\, dh_1 \ ,$$

$$\therefore e^{-h_2} h_2^{\frac{n}{2}} = e^{-h_1} h_1^{\frac{n}{2}} \qquad \ldots (5)$$

(4) and (5) together determine the unbiassed critical region.

5·10. Just as in the case of Fisher's fiducial probability, there

is not particular reason why this method should apply only to sufficient statistics. Theoretically, at least, it can be said to apply to all statistics.

5·11 Method of Confidence Intervals.

The transformation in Pitman's method from the coordinates $\xi_r - a$ to $x_r - t$ (ξ_r and t variable, and x_r and a constants), and the final substitution of a for t, brings that method within the realm of fiducial probability. A method independent of the notion of fiducial probability has been outlined by J. Neyman, [Phil. Trans., 236 A, (1937)], and consists in finding an interval $(\underline{\theta}, \bar{\theta})$ which will cover the true fixed value θ° of θ with any assigned relative frequency. $\underline{\theta}$ and $\bar{\theta}$ are independent of θ and dependent on the random variables.

5·12 The Method.

The random variables are

$$X_1, X_2, \cdots, X_n,$$

and their probability law

$$p(x_1, \cdots, x_n \mid \theta_1, \cdots, \theta_\ell) \equiv p(E \mid \theta_1, \cdots, \theta_\ell).$$

The observed sample, E', is $x_1', \cdots, x_n'$, E' being a fixed point in the n-dimensional sample space, and E

a variable point.

We desire to find a "confidence interval," $(\underline{\theta}, \bar{\theta})$ so that, when θ_1° is given,

$$P\{\underline{\theta}(E) \leq \theta_1^\circ \leq \bar{\theta}(E) \mid \theta_1^\circ\} = \alpha .$$

The points , $x_1, \cdots, x_n, \theta_1$, determine an $(n+1)$ - dimensional space W .

θ_1 = constant determines a hyper-plane $G(\theta_1)$ in W , whilst corresponding to any point E there is a straight line $L(E)$ parallel to the axis of θ_1 .

Similarly we have $G(\theta_1')$ and $L(E')$ corresponding to θ_1' and E' .

$\underline{\theta}(E)$ and $\bar{\theta}(E)$ determine two points $B(E)$ and $C(E)$ on L , with coordinates

$$x_1, \cdots, x_n, \underline{\theta}(E)$$

$$x_1, \cdots, x_n, \bar{\theta}(E)$$

$\{B(E), C(E)\}$ is the image of the confidence interval $\delta(E)$.

If the plane $G(\theta_1')$ cuts the confidence interval $\delta(E')$ at $a(\theta_1', E')$, then $a(\theta_1', E')$ is called a "point of acceptance." As $L(E)$ varies,

the points of acceptance will determine a "region of acceptance" $A(\theta_1', E')$ in $G(\theta_1')$. For all points in $A(\theta_1', E')$,

$$\underline{\theta}(E') \leq \theta_1' \leq \bar{\theta}(E'), \quad \theta_1' \text{ given}.$$

For the remainder of the plane $G(\theta_1')$ lies outside the region of acceptance, the interval $\{B(E'), C(E')\}$ will be wholly above or wholly below the plane.

The fact that $\{\underline{\theta}(E), \bar{\theta}(E)\}$ contains or "covers" θ_1' is denoted by

$$S(E) \subset \theta_1'.$$

If the point E lies in the region of acceptance this is denoted by

$$E \in A'(\theta_1)$$

The following necessary and sufficient conditions hold for a region of acceptance:

(1) $P\{E \in A'(\theta_1) \mid \theta_1\} = \alpha$, whatever $\theta_2, \cdots, \theta_\ell$.

(11) For every E there exists at least one value θ_1' of θ_1 such that

$$E \in A'(\theta_1).$$

(111) If $E \in A'(\theta_1')$ and $E \in A'(\theta_1'')$, then $E \in A'(\theta_1''')$ where $\theta_1' < \theta_1''' < \theta_1''$.

(1V) If $E \in A'(\theta_1)$ for any θ_1, where $\theta_1' < \theta_1 < \theta_1''$, then $E \in A'(\theta_1')$ and $E \in A'(\theta_1'')$.

If these conditions are satisfied by $A'(\theta_1)$, then $\underline{\theta}'(E)$ and $\bar{\theta}'(E)$ are the lower and upper ends of the confidence interval; and

$$P\left\{\underline{\theta}'(E) \leq \theta_1^{\circ} \leq \bar{\theta}'(E)\right\} = \alpha,$$

independently of the values of $\theta_2, \cdots, \theta_\ell$.

5·13· Applying this to the rectangular distribution,

$$\alpha f = \frac{dx}{\theta} \quad , \quad 0 < x < \theta;$$

$$= 0 \quad , \quad \text{otherwise.}$$

We have, for two observations x_1 and x_2

$$p(x_1, x_2 | \theta) = \theta^{-2} \quad , \quad 0 < x_1, x_2 < \theta$$

$$= 0 \quad , \quad \text{otherwise.}$$

Here the sample space W is a plane. p is positive within a region W_+, in this case a square with side θ.

We have to find on G (θ) a region A (θ) satisfying conditions (1) to (1V).

5·14 Example.

The solution is not unique, as the following will show:

(1) Consider $A_1(\theta)$ defined by

$$\theta - \Delta \leq x_1 + x_2 \leq \theta + \Delta$$

$$P\{E \in A_1(\theta) \mid \theta\} = \alpha.$$

Here

$$P\{E \in A_1(\theta) \mid \theta\} = 1 - \left(\frac{\Delta}{\theta}\right)^2 = \alpha,$$

$$\therefore \quad \Delta = \theta(1-\alpha)^{\frac{1}{2}}.$$

$$\therefore \quad \theta\left\{1 - (1-\alpha)^{\frac{1}{2}}\right\} \leq x_1 + x_2 \leq \theta\left\{1 + (1-\alpha)^{\frac{1}{2}}\right\}$$

determines, $A_1(\theta)$; and it may be shown that the regions $A_1(\theta)$ satisfy the four conditions; hence are regions of acceptance.

Hence

$$\underline{\theta}_1(E) = \frac{x_1 + x_2}{1 + (1-\alpha)^{\frac{1}{2}}},$$

$$\bar{\theta}_1(E) = \frac{x_1 + x_2}{1 - (1-\alpha)^{\frac{1}{2}}},$$

$$\delta_1(E) = 2(x_1 + x_2)\frac{\sqrt{1-\alpha}}{\alpha},$$

where $\delta_1(E)$ = length of confidence interval, and measures the accuracy of estimation of θ for a fixed value of α.

(11) Consider $A_2(\theta)$ defined by

$$q\theta \leq L < \theta,$$

where L = larger of the two x's, and $0 < q < 1$.

$$P\{E \in A_2(\theta)|\theta\} = P\{q\theta \leq L < \theta|\theta\} = \alpha.$$

But $P\{L < L'|\theta\}$, where L' = constant $\leq \theta$

$$= \int_0^{L'} p(x_1)\,dx_1 \int_0^{L'} p(x_2)\,dx_2$$

$$= \left(\frac{L'}{\theta}\right)^2$$

$$\therefore P\{q\theta \leq L < \theta|\theta\} = 1 - q^2 = \alpha,$$

$$\therefore q = (1-\alpha)^{\frac{1}{2}},$$

$$\therefore \theta(1-\alpha)^{\frac{1}{2}} \leq L < \theta.$$

This system also satisfies all four conditions, and we see that

$$\underline{\theta}_2(E) = L$$

$$\bar{\theta}_2(E) = L(1-\alpha)^{-\frac{1}{2}}$$

$$\delta_2(E) = \frac{1-(1-\alpha)^{\frac{1}{2}}}{(1-\alpha)^{\frac{1}{2}}} L .$$

Comparing (1) and (11) for $\alpha = \frac{3}{4}$, we get in the respective cases,

(a) $\frac{4}{3}\bar{x} \leq \theta \leq 4\bar{x}$; $\delta_1(E) = \frac{8}{3}\bar{x}$, $\bar{x} = \frac{1}{2}(x_1+x_2)$

(b) $L \leq \theta \leq 2L$; $\delta_2(E) = L$.

If we had two samples (1,1) and (0·1 , 1·9) respectively, (a) would give in both cases

$$\frac{4}{3} \leq \theta \leq 4 ;$$

whilst (b) would give respectively

$$1 \leq \theta \leq 2 \quad , \quad 1{\cdot}9 \leq \theta \leq 3{\cdot}8 .$$

5·15. Remarks on the Solutions.

It will be noticed that the second interval for (b) is entirely contained within the interval for (a). Hence the question arises: what do we mean when we say that each of t hese two intervals has the same probability, $\frac{3}{4}$, of covering the true value, θ° ?

According to the theory of confidence intervals each of the intervals $(\frac{4}{3}, 4)$ and (1.9, 3.8) must be considered "equally reliable" as far as the information provided by the sample (0.1, 1.9) is concerned, because they have each been obtained from the same value of α, namely $\frac{3}{4}$, and each of the respective working rules leads to the conclusion that the true value of θ will, by each method, be covered in 75% of the cases if a long series of samples is taken.

We might be led by intuition to a preference for (b) which yields shorter intervals than (a) in all cases for α fixed = $\frac{3}{4}$ but this choice will have nothing to do with the theory of confidence intervals.

Hence the theory of confidence intervals would appear to have this drawback so far as small samples are concerned, namely that the reduction of the data performed by the estimation of θ from a single small sample is of little use by itself, and only becomes useful insofar as it is "stored up" and combined with the information supplied by a further series of samples.

5·16. Behaviour of Intervals as α Varies.

One further point. Whereas it is true that, as Neyman has pointed out, no matter what the values of x_1 and x_2, if $\alpha = \frac{3}{4}$ the lengths of the intervals δ_2 are less than $\frac{3}{4}$ those of δ_1, yet if we keep our sample fixed and increase α, the opposite is the case. For instance, if we take the

sample $x_1 = 1$, $x_2 = 1$, and give various values to α we get the following table

	$\alpha = \frac{3}{4}$	$\alpha = \frac{8}{9}$	$\alpha = \frac{15}{16}$
δ_1	2·67	1·50	1·07
δ_2	1·00	2·00	3·00

Hence if we set $\alpha = \frac{3}{4}$, (b) gives the shortest confidence interval; but if $\alpha = \frac{8}{9}$, (a) gives the shortest interval for the given sample.

5·17. But the matter is even more serious than would appear from the preceding comments. For if we take our two systems of intervals

(a) $\underline{\theta}_1 = \dfrac{x_1 + x_2}{1 + (1-\alpha)^{\frac{1}{2}}}$,

$\overline{\theta}_1 = \dfrac{x_1 + x_2}{1 - (1-\alpha)^{\frac{1}{2}}}$,

$\delta_1 = 2(x_1 + x_2)\dfrac{\sqrt{1-\alpha}}{\alpha}$;

(b) $\underline{\theta}_2 = L$,

$\overline{\theta}_2 = L(1-\alpha)^{-\frac{1}{2}}$,

$\delta_2 = \dfrac{1 - (1-\alpha)^{\frac{1}{2}}}{(1-\alpha)^{\frac{1}{2}}} L$;

and consider their behaviour in the whole range of possible values of α , (between 0 and 1), x_1 , x_2 being kept constant, we speedily see that system (a), although derived in accordance with the preceding theory, cannot be admitted atall as a suitable system of intervals. For when $\alpha = 1$, that is when we desire an interval that is certain to contain θ , we find $\underline{\theta}_1 = \bar{\theta}_1 = x_1 + x_2$; $\delta_1 = 0$. In other words δ_1 , becomes reduced to a single point. Again, when $\alpha = 0$, $\underline{\theta}_1 = \frac{1}{2}(x_1 + x_2)$ but $\bar{\theta}_1 = \infty$ and δ_1 is infinite. That is we have an infinite interval that cannot contain θ despite the fact that the point ($x_1 + x_2$) lies in this interval. Hence system (a) is quite anomalous.

System (b), on the other hand, behaves correctly insofar as δ_2 increases from a point to an infinite interval as changes from 0 to 1.

Such an examination of the behaviour of δ as α takes various values would then appear to be of extreme importance, and would probably exert a modifying influence on the concept of the shortest system of confidence intervals.

5·18 Simultaneous Estimation.

It is found, however, that the application of the preceding methods to the problem of estimating θ_1 for a distribution containing parameters $\theta_1, \theta_2, \cdots, \theta_\ell$ fails unless

we can find regions $A(\theta_1)$ for which

$$P\{E \in A(\theta_1) \mid \theta_1, \theta_2, \cdots, \theta_\ell\} = \alpha$$

and is independent of $\theta_2, \cdots, \theta_\ell$. Such regions are said to be similar to the sample space with regard to $\theta_2, \cdots, \theta_\ell$, and to have size α.

The solution of the problem will obviously depend upon the existence of a set of statistics $T_1, \cdots, T_s$, ( ) sufficient with respect to $\theta_2, \cdots, \theta_\ell$. If $W(T)$ is the locus $T_1, \cdots, T_s$ all constant, and if $\omega(T)$ is a part of $W(T)$, then, provided $\frac{\partial T}{\partial X}$ is continuous for every T and every X it may be shown that

$$P\{E \in \omega(T) \mid E \in W(T)\}$$

is independent of $\theta_2, \cdots, \theta_\ell$, and is a function of θ_1 only. (The above statement reads: "The probability that E belongs to $\omega(T)$ when E belongs to $W(T)$"

If for θ_1 fixed ($= \theta_1'$)

$$P\{E \in \omega(T) \mid E \in W(T)\} = \alpha, \quad 0 < \alpha < 1,$$

then

$$P\{E \in \omega \mid \theta_1'\} = \alpha,$$

where ω is the n -dimensional region formed by all the $\omega(T)$ for all possible values of $T_1, \cdots, T_s$.

For a worked example, see Neyman (loc. cit.) pp.367-9.

5·19. Shortest System of Confidence Intervals.

When, as generally happens, we find several systems of confidence intervals each covering θ_1^0 with the desired relative frequency, α, we will be led by intuition to regard the shortest system of intervals as being the "best" or "most satisfactory".

If a system S_0 of intervals $\delta_0(E)$ is such that

$$P\{\delta_0(E) \subset \theta_1' \mid \theta_1^0\} \leq P\{\delta(E) \subset \theta_1' \mid \theta_1^0\},$$

where θ_1' is any value of θ_1 other than θ_1^0, and $\delta(E)$ any other system S of intervals corresponding to the same α, then S_0 is called the shortest system of confidence intervals. S_0 will then be less likely to cover incorrect values of θ_1 than S.

If $A_0(\theta_1)$ and $A(\theta_1)$ are the regions leading to S_0 and S, then

$$P\{E \in A_0(\theta_1') \mid \theta_1^0\} \leq P\{E \in A(\theta_1') \mid \theta_1^0\},$$

and

$$P\{E \in A_0(\theta_1') \mid \theta_1'\} = P\{E \in A(\theta_1') \mid \theta_1'\} = \alpha.$$

5·20 One-Sided Estimation.

Shortest systems do not exist for many of the common distributions, for example the normal distribution, and in such cases an alternative method is desirable. Such a method is that of "one-sided estimation."

In many cases we are interested in determining a single (unique) limit which cannot be exceeded, (or, again, a limit below which the parameter cannot fall), except with a certain given relative frequency, $1-\alpha$. These unique upper and lower limits are denoted by $\overline{\overline{\theta}}(E)$ and $\underline{\underline{\theta}}(E)$ respectively.

The conditions involved in such a problem are:

(a) $$P\{E \in A_0(\theta_1) \mid \theta_1\} = \alpha \ ;$$

(b) $$P\{E \in A_0(\theta_1') \mid \theta_1''\} = P\{E \in A(\theta_1') \mid \theta_1''\}.$$

where $A(\theta_1)$ satisfies (a)

(c) $\theta_1' - \theta_1''$ is either always positive or always negative.

For example, to determine $\underline{\underline{\theta}}(E)$:

$$P\{\underline{\underline{\theta}}(E) = \theta_1^\circ \mid \theta_1^\circ\} = \alpha \ ;$$

and this leads to choosing the appropriate regions of acceptance there being an infinity of solutions.

Let θ_1' and θ_1'' be any two other values of θ_1 such that

$$\theta_1' < \theta_1^0 < \theta_1''.$$

$\underline{\underline{\theta}}(E)$ need not now be less than θ_1''. If $\underline{\underline{\theta}}(E)$ is less than θ_1', then $\underline{\underline{\theta}}(E) < \theta_1^0$, and our statement concerning the value of θ_1, based on $\underline{\underline{\theta}}(E)$, will be correct. But it will also be correct if, say, $\underline{\underline{\theta}}(E)$ lies between θ_1' and θ_1^0, for instance if $\underline{\underline{\theta}}(E) = \frac{1}{2} \times (\theta_1' + \theta_1^0)$; and in this case $\underline{\underline{\theta}}(E) > \theta_1'$.

Hence let us make the chance of $\underline{\underline{\theta}}(E)$ falling short of θ_1' a minimum, whatever the value of θ_1', provided $\theta_1' < \theta_1^0$; i.e.

$$P\{\underline{\underline{\theta}}(E) < \theta_1' \mid \theta_1^0\} = \text{minimum}$$

Hence

$$P\{E \in A_0(\theta_1') \mid \theta_1^0\} \leq P\{E \in A \mid \theta_1^0\},$$

where A is any other region such that

$$P\{E \in A \mid \theta_1^0\} = P\{E \in A_0(\theta_1') \mid \theta_1^0\} = \alpha,$$

whatever the value of θ_1', provided $\theta_1^0 > \theta_1'$.

For the unique upper limit, on the other hand, we have

$$P\{E \in A_0(\theta_1') \mid \theta_1^0\} \leq P\{E \in A \mid \theta_1^0\},$$

whatever the value of θ_1', provided $\theta_1^\circ < \theta_1'$.

5·21. Example.

The application of these methods to the normal curve

$$p(E|\theta_1,\theta_2) = \left(\frac{1}{\theta_2\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2}$$

gives

$$\bar{\bar{\theta}}_1(E) = \bar{x} + ts,$$

$$\underline{\underline{\theta}}_1(E) = \bar{x} - ts,$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_i) \quad ; \quad s^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2;$$

and t is found from Fisher's tables for $P = 2(1-\alpha)$.

[See Clopper and Pearson: Biom., 26, (1934)].